Handbook of Regional Growth and Development Theories

Edited by Roberta Capello and Peter Nijkamp

HANDBOOK OF REGIONAL GROWTH AND DEVELOPMENT THEORIES

Handbook of Regional Growth and Development Theories

Edited by

Roberta Capello

Politecnico di Milano, Italy

and

Peter Nijkamp

VU University Amsterdam, the Netherlands

Edward Elgar
Cheltenham, UK • Northampton, MA, USA

Published by
Edward Elgar Publishing Limited
The Lypiatts
15 Lansdown Road
Cheltenham
Glos GL50 2JA
UK

Edward Elgar Publishing, Inc.
William Pratt House
9 Dewey Court
Northampton
Massachusetts 01060
USA

A catalogue record for this book
is available from the British Library

Library of Congress Control Number: 2008017423

PEFC/16-33-111
CATG-PEFC-052
www.pefc.org

ISBN 978 1 84720 506 3 (cased)

Printed and bound in Great Britain by MPG Books Ltd, Bodmin, Cornwall

Contents

Contributors

Maria Abreu is Research Associate at the Centre for Business Research, University of Cambridge, UK and Fellow of the Programme on Regional Innovation, Cambridge–MIT Institute, UK.

T. Taylor Aldridge is the Chief of Staff and a Research Fellow in the Entrepreneurship, Growth and Public Policy division at the Max Planck Institute of Economics, Jena, Germany and a PhD student at the University of Augsburg, Germany.

David B. Audretsch is the Director of the Entrepreneurship, Growth and Public Policy division at the Max Planck Institute of Economics, Jena, Germany, a Scholar-in-Residence at the Ewing Marion Kauffman Foundation and the Ameritech Chair of Economic Development at Indiana University, Bloomington, USA.

Amitrajeet A. Batabyal is Arthur J. Gosnell Professor of Economics at the Rochester Institute of Technology, USA.

Johannes Bröcker is Professor of International and Regional Economics at the University of Kiel, Germany.

Ken J. Button is University Professor and Director of the Center for Transportation Policy, Operations and Logistics at George Mason University, USA.

Roberto Camagni is Professor of Urban Economics at the Politecnico di Milano, Italy, and Past President of the European Regional Science Association.

Roberta Capello is Professor of Regional Economics at the Politecnico di Milano, Italy, and President of the Regional Science Association International.

Jeffrey P. Cohen is Associate Professor of Economics at the University of Hartford, and is President of the Transportation and Public Utilities Group (TPUG) of the Allied Social Sciences Association (ASSA).

Henri L.F. de Groot is Associate Professor at the Department of Spatial Economics at the VU University Amsterdam in the Netherlands.

Daria Denti is Researcher and Consultant at ASTER SCPA, the Consortium for industrial research, technological transfer and innovation of Emilia-Romagna and a PhD awarded by the European University Institute, Italy.

Kieran P. Donaghy is Professor of City and Regional Planning at Cornell University, USA.

Cem Ertur is Professor in Econometrics at the University of Orléans and researcher at the Laboratory of Economics of Orléans, France.

Roberto Ezcurra is Assistant Professor of Economics at the Universidad Pública de Navarra, Spain.

Alessandra Faggian is Reader in Economic Geography in the School of Geography, University of Southampton, UK.

Manfred M. Fischer is Professor of Economic Geography and Dean for the Social Sciences at the Vienna University of Economics and Business Austria, and BA, and Fellow of the Regional Science Association International.

Henk Folmer is Professor of Research Methodology and Spatial Econometrics at the University of Groningen and Professor of General Economics at Wageningen University, the Netherlands.

Peter Friedrich is Extraordinary Professor of Public Economics, University of Tartu, Faculty of Economics and Business Administration, Tartu, Estonia.

Börje Johansson is Professor of Economics at Jönköping International Business School (JIBS), director of the Centre of Excellence for Science and Innovation Studies (CESIS) at the Royal Institute of Technology, Stockholm, Sweden and Past President of the European Regional Science Association.

Charlie Karlsson is Professor of the Economics of Technological Change and Director of the Centre for Innovation Systems, Entrepreneurship and Growth, Jönköping International Business School, Jönköping University, Sweden and Guest Professor of Economics, University West, Trollhättan, Sweden.

T.R. Lakshmanan is Professor and Director, Center for Transportation Studies at Boston University, USA.

Julie Le Gallo is Professor of Economics and Econometrics at the Université de Franche-Comté, France.

Gunther Maier is Associate Professor in Regional Development at Vienna University of Economics and Business, Vienna, Austria, and leads the Research Institute for Spatial and Real Estate Economics at this university.

Philip McCann is Professor of Economics, University of Waikato, New Zealand and Professor of Urban and Regional Economics, University of Reading, UK.

G. Alfredo Minerva is Assistant Professor in Economics at the University of Bologna, Italy.

Ryohei Nakamura is Professor of Urban and Regional Economics Policy at Okayama University, Japan.

Chang Woon Nam is Senior Research Fellow at the Ifo Institute for Economic Research in Munich, Germany.

Peter Nijkamp is Professor in the Department of Spatial Economics at the VU University Amsterdam, the Netherlands.

Jan Oosterhaven is Professor of Spatial Economics at the University of Groningen, the Netherlands, and currently President of the International Input–Output Association.

Gianmarco I.P. Ottaviano is Professor of Economics at the University of Bologna, Italy.

Catherine J. Morrison Paul is Professor in the Department of Agricultural and Resource Economics at the University of California, Davis, USA and member of the Giannini Foundation.

Karen R. Polenske is Professor of Regional Political Economy and Planning and current Fellow of the International Input–Output Association and of the Regional Science Association International, and Past President of the International Input–Output Association.

Jacques Poot is Professor of Population Economics at the University of Waikato in New Zealand.

Piet Rietveld is Professor of Transport Economics at the Faculty of Economics, Vrije Universiteit, Amsterdam, the Netherlands and Past President of NECTAR.

Alistair Robson is a Research Fellow in the UQ Social Research Centre in the Institute of Social Science Research at the University of Queensland, Australia.

Andrés Rodríguez-Pose is Professor of Economic Geography at the London School of Economics, UK.

Maria Savona is Associate Professor in Economics at the Faculty of Economics and Social Sciences, University of Lille 1, France and Visiting Research Fellow at the Cambridge–MIT Institute, University of Cambridge, UK and at SPRU, Science and Technology Policy Research, University of Sussex, UK.

Tung-Kai Shyy is a Research Fellow in the UQ Social Research Centre in the Institute for Social Science Research at the University of Queensland, Australia.

Martijn J. Smit is a PhD student at the Department of Spatial Economics at the VU University Amsterdam in the Netherlands.

Robert J. Stimson is Professor of Geographical Sciences and Director of the Urban and Regional Analysis Program in the UQ Centre for Social Research in the Institute for Social Science Research at the University of Queensland, Australia. He is also Convenor of the Australian Research Council Research Network in Spatially Integrated Social Science, and is the immediate Past President of the Regional Science Association International.

Michaela Trippl is Researcher at Wirtschaftsuniversität Wien, Vienna, Austria.

Jouke van Dijk is Professor of Regional Labour Market Analysis at the Faculty of Spatial Sciences of the University of Groningen, the Netherlands and the Urban and Regional Studies Institute and Editor-in-Chief of *Papers in Regional Science*.

Frank van Oort is Professor in Urban Economics and Spatial Planning at Utrecht University and the Netherlands Institute of Spatial Research in The Hague, the Netherlands.

Introduction: regional growth and development theories in the twenty-first century – recent theoretical advances and future challenges

Roberta Capello and Peter Nijkamp

I.1 The resurgence of regional economics

Regional economics is back on the stage. Regional development is not only an efficiency issue in economic policy, it is also an equity issue due to the fact that economic development normally exhibits a significant degree of spatial variability. Over the past decades this empirical fact has prompted various strands of research literature, in particular the measurement of interregional disparity, the causal explanation for the emergence or persistent presence of spatial variability in economic development, and the impact assessment of policy measures aimed at coping with undesirable spatial inequity conditions. The study of socio-economic processes and inequalities at meso and regional levels positions regions at the core places of policy action and hence warrants intensive conceptual and applied research efforts.

For decades, the unequal distribution of welfare among regions and/or cities has been a source of concern for both policy-makers and researchers. Regional development is about the geography of welfare and its evolution. It has played a central role in such disciplines as economic geography, regional economics, regional science and economic growth theory. The concept is not static in nature, but refers to complex space–time dynamics of regions (or an interdependent set of regions). Changing regional welfare positions are often hard to measure, and in practice we often use gross domestic product (GDP) per capita (or growth thereof) as a statistical approximation (see Stimson et al., 2006). Sometimes alternative or complementary measures are also used, such as per capita consumption, poverty rates, unemployment rates, labour force participation rates or access to public services. These indicators are more social in nature and are often used in United Nations welfare comparisons. An example of a rather popular index in this framework is the Human Development Index which represents the welfare position of regions or nations on a 0–1 scale using quantifiable standardized social data (such as employment, life expectancy or adult literacy) (see for example Cameron, 2005). In all cases, however, spatial disparity indicators show much variability.

Regional disparities may have significant negative socio-economic cost consequences, for instance, because of social welfare transfers, inefficient production systems (for example due to an inefficient allocation of resources) and undesirable social conditions (see Gilles, 1998). Given a neoclassical framework of analysis, these disparities (for example in terms of per capita income) are assumed to vanish in the long run, because of the spatial mobility of production factors which causes at the end an equalization of factor productivity in all regions. Clearly, long-range factors such as education, research and development (R&D) and technology play a critical structural role in this context. In the short run, however, regional disparities may show rather persistent trends (see also Patuelli, 2007).

Disparities can be measured in various relevant categories, such as (un)employment, income, investment, growth, and so on. Clearly, such indicators are not entirely independent, as is, for instance, illustrated in Okun's law, which assumes a relationship between economic output and unemployment (see Okun, 1970; Paldam, 1987). Convergence of regional disparities is clearly a complex phenomenon which refers to the mechanisms through which differences in welfare between regions may vanish (see Armstrong, 1995). In the convergence debate, we observe increasingly more attention for the openness of spatial systems, reflected *inter alia* in trade, labour mobility, commuting, and so on (see for example, Magrini, 2004). In a comparative static sense, convergence may have varying meanings in a discussion on a possible reduction in spatial disparities among regions, in particular (see also Barro and Sala-i-Martin, 1992; Baumol, 1986; Bernard and Durlauf, 1996; Boldrin and Canova, 2001):

- β-convergence: a negative relationship between per capita income growth and the level of per capita income in the initial period (for example, poor regions grow faster than initially rich regions);
- σ-convergence: a decline in the dispersion of per capita income between regions over time.

The convergence hypothesis in neoclassical economics has been widely accepted in the literature, but is critically dependent on two hypotheses (see Cheshire and Carbonaro, 1995; Dewhurst and Mutis-Gaitan, 1995):

- diminishing returns to scale, which means that output growth will be less than proportional with respect to capital growth;
- technological progress will generate benefits that also decrease with its accumulation (that is, diminishing returns).

Many studies have been carried out to estimate the degree of β-convergence and σ-convergence (see for example Barro and Sala-i-Martin, 1991, 1992). The general findings are that the rate of β-convergence is in the order of magnitude of 2 per cent annually, while the degree of σ-convergence tends to decline over time, for both US states and European regions. Clearly, there is still an ongoing debate worldwide on the type of convergence, its speed, its multidimensional conceptualization, and its causal significance in the context of regional policy measures (see for example Fagerberg and Verspagen, 1996; Fingleton, 1999; Galor, 1996). Important research topics in the current literature appear to be: the role of knowledge and entrepreneurship, spatial heterogeneity in locational or socio-cultural conditions, and institutional and physical barriers. An important new topic in the field has become group convergence (or club convergence) (see for example Islam, 2003; Fischer and Stirbock, 2006; Baumont et al., 2003; Chatterji, 1992; Chatterji and Dewhurst, 1996; López-Bazo et al., 1999; Quah, 1996; Rey and Montouri, 1999; Sala-i-Martin, 1996). Thus we may conclude that the research field of spatial disparities is still developing and is prompting fascinating policy issues.

In the light of the previous observations, it is no surprise that over the last decade a resurgence of interest in regional science has taken place, from both theoretical and policy perspectives. This is particularly evident in the case of Europe. One of the main reasons

for such a renewed interest also relates to recent institutional agreements: in May 2004 and in January 2007 the European Union recorded two important historic enlargements, achieving respectively 25 and 27 EU member states. Most of the Eastern European countries joined the European Union, with the consequence of a drastic increase in regional disparities. In May 2004 the enlargement added 5 per cent to the GDP of the EU and 20 per cent to its population; as a consequence, however, the per capita GDP dropped by 12.5 per cent on the day of the enlargement. In January 2007, with the entrance of Romania and Bulgaria, the situation became even worse. Social, economic and demographic disparities call nowadays for sound regional policies.

Clearly, old issues, like regional disparities and convergence, are not the only reasons explaining the resurgence of regional science. Interestingly enough, in recent times, new normative principles in relation to regional development in the European Union have been proposed in official documents; 'territorial cohesion' is quoted in the official EU policy documents as a strategic principle, as strategic as the Lisbon and Gothenburg principles (Luxembourg Presidency, 2005a, 2005b): 'In practical terms territorial cohesion implies: *focusing regional and national territorial development policies* on better exploiting regional potentials and territorial capital – Europe's territorial and cultural diversity; *better positioning of regions in Europe* . . . facilitating their connectivity and territorial integration; and *promoting the coherence of EU policies with a territorial impact*' (p. I). Given the strong attention given by policy-makers to territorial aspects, regional science (and within it, regional economics) has to provide solid theoretical and methodological tools upon which normative policies can be built.

The interest of policy-makers for territorial and regional issues partly explains the resurgence of interest in regional science, and regional economics. Besides policy issues, in the academic arena much interest over the 1990s has also arisen in spatial phenomena: the role of space, highly neglected by mainstream economists, has now become a source of scientific thinking within traditional macroeconomic, international and industrial economic disciplines, giving rise to partly new and partly revisited theories. The degree of convergence and cross-fertilization of ideas between regional economists and the mainstream economists is still an open debate.

Moreover, in a period of globalization like the present one, and the creation of broad single-currency areas, regions (and also nations) must closely concern themselves with the competitiveness of their production systems, because no spontaneous or automatic adjustment mechanism is at work to counterbalance a lack (or an insufficient growth rate) of productivity. Local specificities and local material and non-material assets become strategic elements upon which the competitiveness of regions is based. Theories of regional growth and development need to be able to interpret, more than ever, the way in which regions achieve a role in the international division of labour and, more importantly, the way in which regions can maintain this role over time.

The focus of this volume is to collect the most advanced theories explaining regional growth and local development, with the intention to highlight: (1) the recent advances in theories; (2) the normative potentialities of these theories; (3) the cross-fertilization of ideas among regional economists and mainstream economists.

The aim of this introductory chapter is to summarize the main messages emerging from a package of 25 chapters present in the book, leaving each single chapter to present underlying theories and principles in more detail. Section I.2 will now present the

Table I.1 Main tendencies in theories of regional economics

Tendencies in theories	Regional growth theories	Regional development theories
More realism in theoretical approaches	Endogenous growth determinants A role in growth models of the complex non-linear and interactive behaviours and processes that take place in space Imperfect market conditions in growth models Growth as a long-term competitiveness issue Technological progress as an endogenous factor of growth	Reasons of success and failure of SME cluster areas, local districts, milieux Non-material resources as sources of regional competitiveness An active role of space in knowledge creation
Dynamic rather than static approaches	Evolutionary trajectories of non-linear interdependencies of complex systems	Dynamic rather than static agglomeration economies

Source: Capello (2008).

theoretical progress recently achieved in different parts of the world; section I.3 will then deal with future challenges in this field; while section I.4 will present the main structure of the book.

I.2 Recent theoretical directions

The great number of relatively new and advanced contributions in the fields of regional development and growth theories does not allow for a detailed review on all individual achievements made; in addition, a disaggregated analysis of all novelties would probably not be so stimulating. Our impression is that an attempt to highlight general theoretical trends will turn out to be more fruitful for a debate on present weaknesses and on possible future directions of regional economics. Inevitably, the set of 'tendencies' that follows is both selective and incomplete, primarily reflecting personal views and research interests (Table I.1) (Capello, 2008).

By looking at the theoretical trajectories followed in regional economics, one of the major tendencies which has accompanied the theoretical development in the field is the need for more realism in sometimes rather abstract conceptual approaches, by relaxing most of the glaring unrealistic assumptions of the basic theoretical models, a tendency common also in urban economics (Capello and Nijkamp, 2004). This tendency is justified by the need to broaden the interpretative capacity of the theoretical toolbox in this research field by searching for theories that are better able to reflect the real world.

In regional growth theories, more realism has required the insertion of the complex non-linear and interactive behaviours and processes that take place in space into growth models, and the understanding of regional competitiveness in terms of endogenous factors. The question of whether a region is intrinsically capable of growing as a result of endogenous forces has been a source of debate for decades; industrial specialization, infrastructure endowment, central location, production factor endowment or agglomeration

economies have alternatively been emphasized in the academic arena as driving forces of local economic success.

The decisive step forward in this field has been the focus on economies of scale in production which, together with non-linear transportation costs, are introduced into a (quantitative) interregional growth model; the final spatial distribution of activities critically depends on initial conditions including the starting distribution of activities and the nature of the non-linearities embedded in the activity–transportation interactions, which give rise to multiple equilibria (Krugman, 1991). The additional value of this approach – known as the 'new economic geography' – resides in skilfully modelling the interaction between transportation costs and economies of scale in production, although the determinants of endogenous growth have already long since been emphasized, starting from the Myrdal–Kaldor model (increasing returns, cumulative self-reinforcing growth patterns). The aim to incorporate agglomeration economies – in the form of increasing returns – into elegant models of a strictly macroeconomic nature was made possible by advances in more sophisticated mathematical tools for analysis of the qualitative behaviour of dynamic non-linear systems (bifurcation, catastrophe and chaos theory) together with the advent of formalized economic models which abandoned the hypotheses of constant returns and perfect competition (Fujita and Thisse, 1996, 2002).

These new theoretical advances required a new conceptualization of space, that of a diversified-stylized space (see Capello, this volume). Space is, in these new theories of local growth, a diversified space, since the existence of polarities in space is envisaged where development takes place, diversifying the level and rate of income growth even among areas of the same region. However, it is a stylized space, since polarities are treated as points devoid of any territorial dimension. This approach moves away from the concept of a uniform-abstract space of growth theories developed in the 1950s and 1960s; the label 'uniform' stems from the fact that in these theories supply conditions (factor endowment, sectoral and productive structure) and demand conditions (consumer tastes and preferences) are identical everywhere in the region; abstract, since simplifying assumptions are inserted so as to cope with place-specific conditions (see Capello, 2007a and Chapter 2 in this volume).

In parallel with Krugman's efforts, in the field of endogenous determinants great emphasis has recently been put on knowledge as a driving force to development, and, what is really new, on the endogenous self-reinforcing mechanisms of knowledge creation. Macroeconomic models of endogenous growth, where knowledge is generally embedded in human capital (Romer, 1986; Lucas, 1988), have widely dominated the academic arena in the 1990s. Their main aim was to insert more realism into growth models by relaxing the unrealistic assumption that technological progress is an exogenous process in an economic system; in the new growth theories, instead, technological progress is an endogenous response of economic actors in a competitive environment. More specifically, increasing returns in factor productivity stemming from endogenous factors – such as innovation, scale economies and learning processes – are included in a neoclassical production function, where they offset the effect of the marginal productivity of the individual factors, which the traditional neoclassical approach assumes to be decreasing.

The identification of endogenous determinants of growth was the crucial scientific issue that explained the birth of regional development theories. Development is in fact by definition endogenous. It is fundamentally dependent on a concentrated organization of

the territory, embedded in which is a socio-economic and cultural system whose components determine the success of the local economy: entrepreneurial ability, local production factors (labour and capital), relational skills of local actors generating cumulative knowledge-acquisition and, moreover, a decision-making capacity which enables local economic and social actors to guide the development process, support it when undergoing change and innovation, and enrich it with the external information and knowledge required to harness it to the general process of growth, and to the social, technological and cultural transformation of the world economy. The micro-behavioural nature of these approaches allows a deep understanding of the sources of territorial externalities, of increasing returns in the form of agglomeration economies, at the basis of industrial cluster formation. Within this approach, much emphasis is given to the role of entrepreneurship in regional development (Nijkamp and Stough, 2004).

More realism in the study of clusters and their determinants called for a better understanding of successes and failures of local productive systems, hardly explained in the first theories proposed. Dynamic agglomeration economies – defined as territorial advantages that act on the capacity of firms and regions to innovate – become the centre of most recent theoretical reflections in this field, giving rise to neo-Schumpeterian approaches in regional development. A major debate dominates the academic arena, with the aim to identify the role of space in innovative processes.

In the vast literature created in this field, the endogenous determinants of innovation are increasing returns in the form of dynamic location advantages deriving from:[1] (1) spatial, geographical proximity among firms, which facilitates the exchange of tacit knowledge: this characterizes reflection by economic geographers concerned to explain the concentration of innovative activities; (2) relational proximity among firms, defined as interaction and cooperativeness among local agents, the source of collective learning processes and socialization to the risk of innovation (that is, territorialized relations among subjects operating in geographical and social proximity): this was the approach taken by territorial economists in explaining the dynamic of local systems in terms of local innovative capacity; (3) institutional proximity taking the form of rules, codes and norms of behaviour which facilitate cooperation among actors and therefore the socialization of knowledge and assist economic actors (individual people, firms and local institutions) to develop organizational forms which support interactive learning processes: this aspect was emphasized by more systemic approaches seeking to understand the evolution of complex systems like the innovative system.

A second clear tendency in theoretical developments – typical of regional development and growth theories only – has been the attempts to move towards dynamic approaches. Time matters as well as space in regional science, and this also holds in regional economics. The effort to encapsulate time in spatial analyses has taken place in two different ways, according to two different meanings of time applied in the two fields of analysis: a more traditional chronological time, and time as rhythm of innovative phenomena which occur in the territory which has been applied in regional growth models.

The introduction of a chronological time within spatial analysis is not at all a simple task, since it requires a mathematical and methodological toolbox, only recently available to regional scientists. Theories on non-linear regional dynamics – framed in the context of chaos theory, synergetics theory or predator–prey analysis – may be mentioned here (see Nijkamp and Reggiani, 1999). In growth models, until a few years ago, the large

majority of experiments and applications has taken for granted the existence of linear – and thus regular – growth processes. Linear models are certainly able to generate unstable solutions, but the solutions of such models are restricted to certain regular standard types. Such models may provide approximate replications of short- and medium-run changes, but fail to encapsulate long-term developments characterized by structural shifts of an irregular nature. This limit has recently been overcome with the adoption of non-linear models, which allow for a change in the dynamics of a system generated even by small perturbations in structural forms; structural instability means the possible existence of significant qualitative changes in the behaviour of the system (that is, in the state variables) that are closely connected with bifurcation and catastrophe phenomena that can occur if the parameter values (that is, the control variable) are changing (see Fujita and Thisse, 1996, 2002). The application of non-linear models to the well-known neoclassical and Keynesian models has shown that the deterministic and unique results achieved by the dynamic linear models are no longer guaranteed: interregional income convergence determined by the traditional neoclassical model collapses and opens the way to alternative possible trajectories, and equilibria solutions; non-linear Keynesian Myrdal–Kaldor models substitute the deterministic result of continuous growth or decline with new and opposite development trajectories, after catastrophe phenomena occur (Miyao, 1984, 1987a, 1987b).

Such a theoretical improvement has also been useful in achieving a greater realism of these models, able to incorporate the dynamic interactions between the components of a spatial system. Dynamic interactions are functionally determined by interdependencies between the behaviour of actors and distance frictions. Such spatial interactions may be stable in nature (that is, operating under fixed external conditions) or subject to change as a result of dissipative evolutionary processes in the external world. In the latter case, model parameters become time-dependent, so that non-linear complex dynamics may emerge (see Puu, 1991; Nijkamp and Reggiani, 1993, 1999; Nijkamp, 2006).

In the field of regional development, conceptually speaking a different concept of time has been developed and applied; time à la Bergson–Heidegger is interpreted as duration and a continuous process of creation, characterized by discontinuity, irreversibility, sequentiality and cumulativity. Time has thus been conceived by an important part of regional studies as the pace of learning, innovation and creation processes. Local clusters (and industrial districts) are by definition the loci where learning and cumulative learning processes take place; the identification of the sources and of the endogenous determinants of such processes, besides simple physical proximity, represents a great challenge for regional economists. Knowledge spillovers, collective learning, learning regions (or learning space) and knowledge-based regions are all theories that embrace the most advanced perspectives in this direction. In these theoretical approaches, therefore, innovation has become the critical survival factor in a competitive space-economy and determines the direction and pace of regional development (Nijkamp and Abreu, 2008).

I.3 Future challenges

Fascinating new theoretical challenges are nowadays faced by regional scientists, and have to be addressed. A first challenge is proposed by the attempt to obtain advantages by a future convergence in different theoretical approaches, a convergence only partially obtained by the new regional growth theories. New growth theories make a commendable

effort to include space in strictly economic models. Also to be commended is the implicit merging in their theoretical structure of the various conceptions of space put forward over the years: the merging, that is, of the physical-metric space represented by transport costs with the diversified space which assumes the hypothesis of the existence of certain territorial polarities where growth cumulates. However, the new economic geography is still unable to combine the economic laws and mechanisms that explain growth with territorial factors springing from the intrinsic relationality present at local level. An approach that did so would represent the maximum of cross-fertilization among location theory, development theory and macroeconomic growth theory; a synthesis which would bring out the territorial micro-foundations of macroeconomic growth models (Capello, 2007a and Chapter 2 in this volume).

Still needed, therefore, is a convincing 'model' which comprises the micro-territorial, micro-behavioural and intangible elements of the development process. Required for this purpose are definitions of patterns, indicators and analytical solutions to be incorporated into formalized models necessarily more abstract and synthetic in terms of their explanatory variables; variables besides the cost of transport, which cancels the territory's role in the development process. A move in this direction is the quantitative sociology that embraces the paradigm of methodological individualism and seeks to 'measure' the social capital of local communities. It is obviously necessary to bring out territorial specificities within a macroeconomic model. Or in other words, it is necessary to demonstrate the territorial micro-foundations of macroeconomic growth models.

Another challenge faced by regional scientists is the exploitation deriving from cross-fertilization of interdisciplinary approaches, a limit already underlined during the 1990s, during the reflections on the health of regional science. Since this problem has been underlined, hardly any signs of recovery have been identified, and we feel that the situation has become even more problematic.[2] This pessimistic interpretation is based on some clear tendencies encountered in some recent theoretical developments, where some wide fields of unexplored interdisciplinarity still exist and no tendency to fill them seems to show up.

Some examples are useful in this respect. The theory on 'social capital' developed by quantitative sociology is an example in this respect: the concept could take advantage from and provide advantage to all reflections on local synergies and milieu effects developed by regional and urban economists, and by the strategic planning studies in the field of urban planning. The reflections in the field of knowledge spillovers developed by industrial economists could take advantage from the concepts of collective learning and relational proximity of regional scientists, in which the endogenous spatial development patterns of knowledge are not left to simple probabilistic contacts, but explained through territorial processes (Camagni and Capello, 2002). Last but not least, the theoretical reflections characterizing the 'new economic geography' seem to be the result of the skilful effort of a group of mainstream economists, driven however by a somehow unexplainable attitude to deny the importance of well-known spatial concepts (that is, technological spatial externalities), or to (re)invent important spatial concepts (that is, cumulative self-reinforcing processes of growth; transportation costs versus agglomeration economies in location choices). The inevitable consequence of this attitude is to mix the important and undeniable steps forward made by the 'new economic geography' school with already well-known knowledge in the field of regional science.

Some risks of disciplinary barriers and of closeness to interdisciplinary views on strategic problems are still there. They are the result of a regional scientists' narrow perspective, as mentioned by Bailly and Coffey (1994), but also on some idiosyncratic approaches of mainstream disciplines towards a clearly multidisciplinary science like regional science. Especially in the case of economics, we hope that after the (re)discovered interest of mainstream economists in space, and in spatial phenomena, the attitude towards regional science will change in favour of a more cooperative attitude and pronounced interest.

Related to the interdisciplinary challenge, a last important remark is worth mentioning. An interdisciplinary approach should lead scientists to explore new frontiers and achieve new interpretative analytical frameworks. The tendency shown in this respect is a different one, more inclined to exploit passively the new ideas suggested by complementary disciplines. A case in this respect that is worth mentioning is the enthusiastic way in which regional scientists accepted the spatial spillover theory as a theory adding a new interpretation to the explanation of the role of space as a knowledge transition.

Instead, a critical approach to this theory shows that under certain respects this theory has made some steps backwards in the interpretation of space in spatial knowledge creation. Space is purely geographical, a physical distance among actors, a pure physical container of spillover effects which come about – according to the epidemiological logic adopted – simply as a result of physical contact among actors. Important consequences ensue from this interpretation of space. Firstly, this view is unable to explain the processes by which knowledge spreads at local level, given that it only envisages the probability of contact among potential innovators as the source of spatial diffusion. Secondly, it concerns itself only with the diffusion of innovation, not with the processes of knowledge creation. It thus imposes the same limitations as did Hägerstrand's pioneering model in regard to the spatial diffusion of innovation: the diffusion of knowledge means adoption, and adoption means more innovation and better performance. Thus ignored, however, is the most crucial aspect of the innovation process: how people (or the context) actually learn. This calls for a more thorough and innovative investigation of cognitive processes in a regional context (Capello, 2008). This is the aspect of overriding interest not only for scholars but also, and especially, for policy-makers, should they wish to explore the possibilities of normative action to promote local development.

I.4 Structure and content of the volume

The volume is organized in five parts, reflecting the new theoretical directions emphasized in previous sections. Part I is built around the new concepts of space and growth that are nowadays at the basis of regional growth and development theories. After a historical perspective provided by the first chapter on the development of theoretical approaches, Chapter 2 introduces the new concepts of growth and space, highlighting the major steps forward made in introducing space in regional growth models, and in defining growth. Chapter 3 deals with the interpretative capacities of theories on the spatial distribution of regional growth, underlining the achievements made in the neoclassical approach to regional growth by moving from constant to increasing returns to scale, thanks to the introduction of externalities into a general equilibrium model to explain long-term growth processes. This chapter introduces the new economic geography (NEG) theory presented in Chapters 4 and 5. The purpose of Chapter 4 is to provide a selective survey of different aspects of the relationship between trade and regional growth that existing

theories of trade, agglomeration and fragmentation can help us to understand, and to indicate where the frontiers of research lie. Chapter 5 describes a simple theoretical framework to study the impact of infrastructure on economic growth and regional imbalances within the framework of NEG models with endogenous growth and free capital mobility.

Part II is devoted to advances in regional development theories, with a particular emphasis on production factors endowment. The first chapter of Part II, Chapter 6, recalls causes and effects of agglomeration economies, and reviews systematically the ways in which causes and effects of increasing returns due to the density of manufacturing activities over space can be measured through a production function approach. Chapter 7 is devoted to the presentation of a new concept, that of territorial capital, which, strangely enough has only recently made its appearance, and has done so outside a strictly scientific context; as the author mentions, by this term, material and non-material elements characterizing a local area are meant, which define its local competitive capacity. Within these elements, cognitive aspects are also analysed; the way economic agents perceive economic reality, are receptive to external stimuli, can react creatively, and are able to cooperate and work synergetically becomes a strategic aspect. Local competitiveness is interpreted as residing in local trust and a sense of belonging rather than in pure availability of capital; in creativity rather than in the pure presence of skilled labour; in connectivity and relationality more than in pure accessibility; in local identity besides local efficiency and quality of life. Chapter 8 focuses on an important intangible asset explaining local competitiveness, that of human capital, which is one of the most important elements defining the territorial capital of a region. The chapter recalls that the links between human capital and national economic development may not necessarily be the same as those between human capital and regional economic development. Two quite distinct sets of human capital impacts on regions exist, the first of which mirrors the national impact, while the second differs markedly from the national impact. The human capital in a region in fact has an impact on the aggregate productivity in the economy, via the externalities associated with it, as at the national level. However, rather differently to national economies, human capital in a region can also result in a major spatial reallocation of factors. Chapter 9 looks at regional impacts of infrastructure supply; the chapter deals with the considerable uncertainty that often exists in relation to the regional economic effects of infrastructure supply, and the measurement of impacts in terms of both a productivity orientation and a welfare orientation, going from computable general equilibrium (CGE) models to a method that is much less demanding in terms of data as well as computational complexity, but still theoretically well founded and closely related to a familiar approach in regional science: gravity analysis. Chapter 10 offers a review on modern entrepreneurship analysis, against the background of regional development. After a conceptual discussion on the importance and the measurement of entrepreneurship, the contribution discusses critical success factors and key determinants of entrepreneurship. Next, much focus is laid on the geography of entrepreneurship, while due attention is also paid to the relevance of networks for modern entrepreneurship. The chapter concludes with some retrospective and forward-looking remarks.

Part III is devoted to advances in local development theories with a dynamic approach, where time is interpreted as the rhythm of innovative phenomena which occur in the territory which has been applied in regional growth models. Knowledge creation and

diffusion processes over space have become of primary interest in a knowledge-based society. Chapter 11 opens the debate on the importance of knowledge for regional competitiveness. The chapter explains why the emergence of knowledge as a source of comparative advantage has rendered a shift in the organization of economic activity at both the spatial and enterprise levels. This chapter uses the lens provided by the knowledge spillover theory of entrepreneurship to integrate the organization of enterprise with that of economic activity in geographic space. Chapter 12 provides a review of the most influential and path-breaking works that have tried to respond to the most important issues related to R&D expenditure, namely the effects of publicly funded R&D on industrial productivity growth. The main conclusion is that there is no general policy advice on how to deal optimally with R&D. The variety of proper fiscal tools depends heavily on the menu of R&D spillovers that are influencing the economy. Positive spillovers call for public support, but it may also be the case that R&D exerts negative externality effects. Chapter 13 examines models depicting and explaining the role of knowledge in regional development and provides an assessment of empirical studies of how knowledge affects growth and development in functional regions. In this endeavour, it is crucial to understand those factors that make knowledge spatially sticky and knowledge-production capacity trapped. It is equally important to explain the conditions for knowledge flows and diffusion. The presentation also widens the view by extending the analysis of knowledge creation to include aspects of creativity. In a part devoted to knowledge, innovation and regional development a review of modern theories and approaches on the role of innovation on regional development is important. Chapter 14 revisits the central part of this virtuous circle, namely the Marshall–Arrow–Romer externalities (specialization), Jacobs externalities (diversity) and Porter externalities (competition) that have provided alternative explanations for innovation and regional (urban) growth. The aim of the chapter is to explain variation in estimation results using study characteristics by means of ordered probit analysis. The evidence in the literature on the role of the specific externalities is rather mixed, although for each type of externality we can identify how various aspects of primary study design influence the outcomes. The chapter evaluates the statistical robustness of evidence for such externalities presented in 31 scientific articles, all building on the seminal work of Glaeser et al. (1992). Chapter 15 deals with the increasing amount of research now being conducted on topics at the interface of regional growth and sustainable development. Specifically, the chapter focuses on five key issues and these issues are: (1) regional economic development; (2) natural resources; (3) environmental regulation; (4) geographic information systems; and (5) regional climate change. The review is both retrospective and forward-looking, by discussing what has been achieved thus far and the likely future directions of research on regional growth and the sustainable development.

Part IV deals with the most advanced methods for measuring regional growth and development. Chapter 16 focuses on the measurement of economic agglomeration in the context of the clustering of regional economic activity. In the chapter various agglomeration measures that have been proposed in the literature are first discussed, in order to provide alternative methodologies for the direct measurement of agglomeration. The estimation of the determinants or sources of agglomeration, and the resulting agglomeration economies or productivity effects of agglomeration, both of which involve methods of indirect measurement, are then discussed. Some topics that will be important to address in future studies of agglomeration economies are also recalled. Chapter 17 provides an

overview of the main developments in the measurement of the regional divide, discussing several methodological issues that have arisen since the first attempts to quantify the magnitude of spatial disparities were made. The chapter highlights the implications of the choice of different methods for the perception of the dimension and evolution of regional disparities and illustrates these empirically by resorting to the case of the EU-15 during the period 1980–2002. Chapter 18 first provides an overview discussion of endogenous growth factors. It then proposes a measure of regional endogenous change which is readily calculable from secondary analysis of regional employment data available in the national census. The regional or differential/regional shift component derived from shift-share analysis of employment change over time is proposed as a viable proxy measure as a dependent variable in an endogenous growth model. A series of independent variables which may also be derived from census data are specified in the model as factors likely to explain spatial variability in regional performance on that dependent variable. Those variables are taken as reflecting the types of factors that are proposed in the regional economic development literature as potentially influencing endogenous growth. The results derived from the application of the model across non-metropolitan regions in the state of Queensland, Australia, are presented. The chapter concludes with some thoughts on the emergence of a new paradigm for regional economic development analysis and planning. Chapter 19 deals with spatial econometric technics, and in particular with spatial heterogeneity. This phenomenon can be observed at several spatial scales: behaviours and economic phenomena are not similar in the centre and in the periphery of a city, in an urban region and in a rural region, in the 'West' of the enlarged European Union and in the 'East', and so on. Spatial heterogeneity is one of the two spatial effects analysed by the field of spatial econometrics. This effect operates through the specification of the reaction of the variable of interest to explanatory variables or the specification of its variance. The chapter first presents the main econometric specifications capturing spatial heterogeneity. Here, we focus on structural instability, as well as on specific forms of heteroskedasticity. Secondly, it examines how these specifications can be extended to allow further for spatial autocorrelation in a model of heterogeneous reaction as well as interaction. Heterogeneity can also be modelled using spatial panel data models. Chapter 20 presents computable general equilibrium (CGE) modelling; this is an approach to applied economic analysis in which theories of economy-wide market behavior are used to impose structure in numerical thought experiments concerning matters of trade and development – and related policies – where the relative unavailability of data or the complexity of a theoretical model's specification poses problems for a more traditional analytical or econometric modelling approach. Over the 1980s and 1990s, CGE modelling has developed extensively and has become a stock in trade of regional scientists in particular. More recently, CGE models have taken on an explicit spatial orientation as the focus of modelling exercises has turned to analysis of location-specific impacts of unplanned events and planned industrial, infrastructural, environmental or other types of regional policies. Spatial CGEs have been employed by researchers at various scales of spatial and temporal resolution to examine a wide variety of phenomena. Owing to the paucity of spatial time series, spatial CGEs (SCGEs) provide logical frameworks within which a broad spectrum of spatial economic issues may be analysed. This chapter surveys a representative sample of studies in the recent literature in SCGE modelling and discusses new directions in which SCGE modelling might be taken. Last but not least, Chapter 21 reviews

the basic theory of input–output and socio-economic accounting in terms of some of the significant methodological debates that occur. Although not all developments are region-specific, the chapter covers them because regional analysts are beginning to adopt these theoretical advancements in their work. For the applications, the chapter restricts its review to regional and multi-regional impact analyses and the development of computer programming packages that help analysts to conduct such studies quickly.

Finally, Part V is devoted to regional policy issues. The first chapter of this part (Chapter 22) opens the discussion on regional policy issues by highlighting the role of institutions in shaping economic development; institutions are now seen as comprising of a set of formal and informal rules, including the conditions of their enforcement, following the new institutional economics. The chapter focuses on the different institutional mechanisms that allow for the coordination of regional economic activities in modern capitalistic economies. It inquires into the logic and functions economic institutions follow. It also traces the emergence of regional institutions in the light of the coherence of the institutional attributes with the structural conditions prevailing in a regional economy, and examines the subsequent development and persistence of institutions. While recognizing that each economic institution (the market, firm, state, networks, associations, and so on) has strengths and weaknesses, the chapter underlines that the preferable approach is not to favour one institution but to combine them according to objectives, resources, and the attributes of the goods and services. The theoretical perspective is broadened in recognizing that the operations of regional economic institutions are constrained by the social context in which they are embedded. Chapter 23 proposes a review of regional policy issues and most of the dilemmas related to the implementation of regional development policies. First, there is the dilemma of 'place prosperity versus people prosperity'. At first instance, a direct targeting of individual inequities by means of, for instance, income support seems the preferred strategy. However, 'place prosperity' may still be needed as an independent goal alongside 'people prosperity', as pursuing only the latter may have unwanted indirect effects. The second dilemma regards the issue of 'interregional equity versus national efficiency'. In order to provide a foundation for the discussion of these and other dilemmas and to understand the logic behind regional policy measures, the chapter discusses several theories that underpin the choice between different regional policy strategies, underlining that there does not exist a one-to-one correspondence between theories and instruments, because theories partly overlap and instruments can sometimes be based on more than one theory. Chapter 24 deals with regional disparities in growth and income levels that represent an important challenge for policy-makers in less developed countries, particularly in the context of increasing globalization. A large number of recent empirical contributions have analysed the extent to which developing countries are able to benefit from trade liberalization and other economic reform policies. However, only a few of these contributions are devoted to the impact of these policies on regional income disparities. This chapter reviews the empirical literature on the regional policies in less developed countries, with an illustration based on the case of India. The review shows that regional policies can complement or counteract the effects of national policies, with the effectiveness of specific regional policies depending on the degree of decentralization of the policy-making process, the extent of sectoral specialization across regions and the degree of regional variation in initial endowments of physical and social infrastructure. To end up with, Chapter 25 is

concerned with the fact that theories aimed at investigating and examining development refer mostly to growth, while economic decline and those factors restricting economic development have not been examined exclusively. The chapter also investigates the implications for development policy-making of the lack of a theoretical approach to economic decline. In some European countries and regions debate has been going on about suitable economic and social policy measures to prevent the decline process resulting from population decrease, for instance. Can decline be overcome? Should policy measures be more strongly directed to decline in order to minimize economic losses caused by decline?

Notes

1. For a literature on spatial spillovers see Anselin et al. (1997, 2000), Audretsch and Feldman (1996), Aydalot (1986), De Groot et al. (2001), Feldman (1994), Feldman and Audretsch (1999), Jaffe (1989), Jaffe et al. (1993), Maier and Sedlacek (2005), on collective learning see Camagni (1991), Capello (1999, 2001), Crevoisier and Camagni (2000), Maillat et al. (1993), Rallet (1993), Rallet and Torre (1995), Ratti et al. (1997), Bellet et al. (1993), on learning regions see Lundvall (1992), Lundvall and Johnson (1994), Maskell and Malmberg (1999), on knowledge-based regions see Malecki (2000), Florida (1995), Nijkamp and Stough (2004), Simmie (2001).
2. On the debate on regional science development, see, among others, Bailly (1992), Bailly and Coffey (1994), Bolton and Jensen (1995), Funck (1991), van Geenhuizen and Nijkamp (1996), Isserman (1993, 1995), Quigley (2001). On a recent debate, see the contributions of the Round Table held in Volos during the ERSA Conference, edited by Coccossis and Nijkamp (2007).

References

Anselin, L., A. Varga and Z. Acs (1997), 'Local geographic spillovers between university research and high technology innovations', *Journal of Urban Economics*, **42**, 422–48.
Anselin, L., A. Varga and Z. Acs (2000), 'Geographic and sectoral characteristics of academic knowledge externalities', *Papers in Regional Science*, **79** (4), 435–43.
Armstrong, H.W. (1995), 'Convergence among the regions of the European Union, 1950–1990', *Papers in Regional Science*, **74** (2), 143–52.
Audretsch, D. and M. Feldman (1996), 'R&D spillovers and the geography of innovation and production', *American Economic Review*, **86** (3), 630–40.
Aydalot, Ph. (eds) (1986), *Milieux Innovateurs en Europe*, Paris: GREMI.
Bailly, A. (1992), 'Representation et analyse des territoires: une epistemologie de la science regionale', in P.H. Derycke (ed.), *Espace et dynamique territoriale*, Paris: Economica.
Bailly, H. and W. Coffey (1994), 'Regional science in crisis: a plea for a more open and relevant approach', *Papers in Regional Science*, **73** (1), 3–14.
Barro, R.J. and X. Sala-i-Martin (1991), 'Convergence across states and regions', *Brookings Papers on Economic Activity*, **1**, 107–82.
Barro, R.J. and X. Sala-i-Martin (1992), 'Convergence', *Journal of Political Economics*, **100** (2), 223–51.
Baumol, W.J. (1986), 'Productivity growth, convergence, and welfare: what the long-run data show', *American Economic Review*, **76** (5), 1072–85.
Baumont, C., C. Ertur and J. LeGallo (2003), 'Spatial convergence clubs and the European regional growth process, 1980–1995', in B. Fingleton (ed.), *European Regional Growth*, Berlin: Springer-Verlag, pp. 131–58.
Bellet, M., G. Colletis and Y. Lung (1993), 'Introduction au numéro spécial sur economie et proximité', *Revue d'Economie Régionale et Urbaine*, **3**, 357–61.
Bernard, A.B. and N. Durlauf (1996), 'Interpreting tests of the convergence hypothesis', *Journal of Economics*, **71** (1–2), 161–74.
Boldrin, M. and F. Canova (2001), 'Europe's regions: income disparities and regional policy', *Economic Policy*, **16**, 207–53.
Bolton, R. and R.C. Jensen (1995), 'Regional science and regional practice', *International Regional Science Review*, **18** (2), 133–45.
Camagni, R. (1991), 'Local milieu, uncertainty and innovation networks: towards a new dynamic theory of economic space', in R. Camagni (ed.), *Innovation Networks: Spatial Perspectives*, London: Belhaven-Pinter, pp. 121–44.
Camagni, R. and R. Capello (2002), 'Apprendimento collettivo, innovazione e contesto locale', in R. Camagni and R. Capello (eds), *Apprendimento Collettivo e Competitività Territoriale*, Milan: Franco Angeli, pp. 11–26.
Cameron, R. (2005), 'Spatial economic analysis', *Journal of Development Perspectives*, **1** (1), 146–63.

Capello, R. (1999), 'Spatial Transfer of Knowledge in high-technology milieux: learning vs. collective learning processes', *Regional Studies*, **33** (4), 353–65.
Capello, R. (2001), 'Urban innovation and collective learning. theory and evidence from five metropolitan cities in Europe', in M.M. Fischer and J. Froehlich (eds), *Knowledge, Complexity and Innovation Systems*, Berlin, Heidelberg and New York: Springer, pp. 181–208.
Capello, R. (2007a), *Regional Economics*, London: Routledge.
Capello, R. (2007b), 'A forecasting territorial model of regional growth: the MASST model', *Annals of Regional Science*, **41** (4), 753–87**.**
Capello, R. (2008), 'Regional economics in its fifties: recent theoretical directions and future challenges', *Annals of Regional Science*, **42** (4), 747–67.
Capello, R. (forthcoming), 'Spatial spillovers and regional growth: a cognitive approach', *European Planning Studies*.
Capello, R. and P. Nijkamp (2004), 'The theoretical and methodological toolbox of urban economics: from and towards where?' in R. Capello and P. Nijkamp (eds), *Urban Dynamics and Growth: Advances in Urban Economics*, Amsterdam: Elsevier, pp. 1–30.
Capello, R., R. Camagni, U. Fratesi and B. Chizzolini (2008), *Modelling Regional Scenarios for an Enlarged Europe*, Berlin: Springer Verlag.
Chatterji, M. (1992), 'Convergence clubs and endogenous growth', *Oxford Review of Economic Policy*, **8** (4), 57–69.
Chatterji, M. and J.H.L. Dewhurst (1996), 'Convergence clubs and relative economic performance in Great Britain: 1977–1991', *Regional Studies*, **30** (1), 31–40.
Cheshire, P. and G. Carbonaro (1995), 'Convergence–divergence in regional growth rates: an empty black box?' in H.W. Armstrong and R.W. Vickerman (eds), *Convergence and Divergence among European Regions*, London: Pion, pp. 89–111.
Coccossis, H. and P. Nijkamp (eds) (2007), 'Challenges in regional science', *Italian Journal of Regional Science*, **6** (2), 137–74.
Crevoisier, O. and R. Camagni (eds) (2000), *Les Milieux Urbains: Innovation, Systèmes de Production et Ancrage*, Neuchâtel: EDES.
De Groot, H., P. Nijkamp and Z. Acs (2001), 'Knowledge spill-overs, innovation and regional development', special issue, *Papers in Regional Science*, **80** (3).
Dewhurst, J.H.L. and H. Mutis-Gaitan (1995), 'Varying speeds of regional DGP per capita convergence in the European Union, 1981–91', in H.W. Armstong and R.W. Vickerman (eds), *Convergence and Divergence among European Regions*, London: Pion, pp. 22–39.
Fagerberg, J. and B. Verspagen (1996), 'Heading for convergence? Regional growth in Europe reconsidered', *Journal of Common Market Studies*, **34** (3), 431–48.
Feldman, M. (1994), *The Geography of Innovation*, Boston, MA: Kluwer Academic.
Feldman, M. and D. Audretsch (1999), 'Innovation in cities: science-based diversity, specialisation and localised competition', *European Economic Review*, **43**, 409–29.
Fingleton, B. (1999), 'Estimates of time to economic convergence: an analysis of regions of the European Union', *International Regional Science Review*, **22** (1), 5–34.
Fischer, M. and C. Stirbock (2006), 'Pan-European regional income growth and club-convergence', *Annals of Regional Science*, **40**, 693–721.
Florida, R. (1995), 'Towards the learning region', *Futures*, **27** (5), 527–36.
Fujita, M. and J.-F. Thisse (1996), 'Economics of agglomeration', *Journal of the Japanese and International Economies*, **10**, 339–78.
Fujita, M. and J.-F. Thisse (2002), *Economics of Agglomeration: Cities, Industrial Location and Regional Growth*, Cambridge: Cambridge University Press.
Funck, R. (1991), 'Regional science in transition', *Papers in Regional Science*, **70**, 1–8.
Galor, O. (1996), 'Convergence? Inferences from theoretical models', *Economic Journal*, **106**, 1056–69.
Geenhuizen, M. van and P. Nijkamp (1996), 'Progress in regional science: a European perspective', *International Regional Science Review*, **19** (3), 223–46.
Gilles, S.-P. (1998), 'The political consequence of unemployment', Working Paper 343, Department of Economics, Universitat Pompeu, Fabra.
Glaeser, E.L., H.D. Kallal, J.A. Scheinkman and A. Schleifer (1992), 'Growth in cities', *Journal of Political Economy*, **100**, 1126–52.
Islam, N. (2003), 'What have we learnt from the convergence debate?', *Journal of Economic Surveys*, **17** (3), 309–62.
Isserman, A. M. (1993), 'Lost in space? On the history, status and future of regional science', *Review of Regional Studies*, **23**, 1–50.
Isserman, A.M. (1995), 'The history, status and future of regional science: an American perspective', *International Regional Science Review*, **17** (3), 249–96.
Jaffe, A. (1989), 'Real effects of academic research', *American Economic Review*, **79**, 957–70.

Jaffe, A., M. Trajtenberg and R. Henderson (1993), 'Geographic localisation of knowledge spillovers as evidenced by patent citations', *Quarterly Journal of Economics*, **63**, 577–98.
Krugman, P. (1991), *Geography and Trade*, Cambridge, MA: MIT Press.
López-Bazo, E., E. Vayá, A. Mora and J. Suriñach (1999), 'Regional economic dynamics and convergence in the European Union', *Annals of Regional Science*, **33** (3), 343–70.
Lucas, R. (1988), 'On the mechanics of economic development', *Journal of Monetary Economics*, **22**, 3–42.
Lundvall, B.-A. (1992), 'Introduction', in B.-A. Lundvall (ed.), *National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning*, London: Pinter Publishers, pp. 1–19.
Lundvall, B.-A. and B. Johnson (1994), 'The learning economy', *Journal of Industry Studies*, **1**, 23–42.
Luxembourg Presidency (2005a), 'Conclusion of the informal EU ministerial meeting on territorial cohesion', 20–21 May, http://www.eu2005.lu/en/actualites/documents_travail/2005/05/20regio/Min_DOC_2_MinConcl_fin.pdf.
Luxembourg Presidency (2005b), 'Scoping document and summary of political messages for an assessment of the territorial state and perspectives of the European Union. Towards a stronger European territorial cohesion in the light of the Lisbon and Gothembourg ambitions', May, http://www.eu2005.lu/en/actualites/documents_travail/2005/05/20regio/Min_DOC_1_fin.pdf.
Magrini, S. (2004), 'Regional (di)convergence', in V. Henderson and F.J. Thisse (eds), *Handbook of Regional and Urban Economics*, Vol. 4, Amsterdam: North Holland, pp. 2741–96.
Maier, G. and S. Sedlacek (eds) (2005), *Spillovers and Innovations: Space, Environment and the Economy*, Vienna: Springer-Verlag.
Maillat, D., M. Quévit and L. Senn (eds) (1993), *Réseaux d'Innovation et Milieux Innovateurs: un Pari pour le Développement Régional*, Neuchâtel: EDES.
Malecki, E. (2000), 'Creating and sustaining competitiveness', in J.R. Bryson, P.W. Daniels, N. Henry and J. Pollard (eds), *Knowledge, Space, Economy*, London: Routledge, pp. 103–19.
Maskell, P. and A. Malmberg (1999), 'Localised learning and industrial competitiveness', *Cambridge Journal of Economics*, **23**, 167–85.
Miyao, T. (1984), 'Dynamic models of urban growth and decay: a survey and extensions', paper presented to the second World Conference of Arts and Sciences, Rotterdam, 4–15 June.
Miyao, T. (1987a), 'Dynamic urban models', in E. Mills (ed.), *Urban Economics: Handbook of Regional and Urban Economics*, Vol. 2, Amsterdam: North-Holland, pp. 877–925.
Miyao, T. (1987b), 'Urban growth and dynamics', in T. Miyao and Y. Kanemoto (eds), *Urban Dynamics and Urban Externalities*, Chur, Switzerland and New York: Harwood Academic Publishers, pp. 1–41.
Nijkamp, P. (2006), 'Ceteris paribus: spatial complexity and spatial equilibrium. An interpretative perspective', mimeo.
Nijkamp, P. and M. Abreu (2008), 'Regional development theory', mimeo.
Nijkamp, P. and A. Reggiani (eds) (1993), *Nonlinear Evolution of Spatial Economic Systems*, Berlin: Springer-Verlag.
Nijkamp, P. and A. Reggiani (1999), *The Economics of Complex Spatial Systems*, Amsterdam: Elsevier.
Nijkamp, P. and R. Stough (eds) (2004), *Entrepreneurship and Regional Economic Development*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.
Okun, A.M. (1970), *The Political Economic of Prosperity*, Washington, DC: Brookings Institute.
Paldam, M. (1987), 'How much does one percent of growth change the unemployment rate?', *European Economic Review*, **31**, 306–13.
Patuelli, R. (2007), 'Regional labour markets in Germany', PhD dissertation, Free University, Amsterdam.
Puu, T. (1991), *Non-linear Economic Dynamics*, Berlin: Springer-Verlag.
Quah, D.T. (1996), 'Empirics for economic growth and convergence', *European Economic Review*, **40** (6), 1353–75.
Quigley, J.M. (2001), 'The renaissance in regional research', *Annals of Regional Science*, **35** (2), 167–78.
Rallet, A. (1993), 'Choix de proximité et processus d'innovation technologique', *Revue d'Economie Régionale et Urbaine*, **3**, 365–86.
Rallet, A. and A. Torre (eds) (1995), *Economie Industrielle et Economie Spatiale,* Paris: Economica.
Ratti, R., A. Bramanti and R. Gordon (eds) (1997), *The Dynamics of Innovative Regions*, Aldershot: Ashgate.
Rey, S.J. and B.D. Mantouri (1999), 'US regional income convergence: a spatial econometric perspective', *Regional Studies*, **33** (2), pp. 143–56.
Romer, P. (1986), 'Increasing returns and long-run growth', *Journal of Political Economy*, **94** (5), 1002–37.
Sala-i-Martin, X. (2006), 'The classical approach to convergence analysis', *Economic Journal*, **106**, 343–70.
Simmie, J. (ed.) (2001), *Innovative Cities*, London: Spon, pp. 95–128.
Stimson, R., R. Stough and H. Roberts (2006), *Regional Economic Development*, Berlin: Springer-Verlag.

PART I

GROWTH THEORIES AND SPACE

1 Theories of agglomeration and regional economic growth: a historical review

Philip McCann and Frank van Oort

1.1 Introduction

Some observers have argued that the modern treatment of agglomeration economies and regional growth really represents a rediscovery by economists of well-rehearsed concepts and ideas with a long pedigree in economic geography. As such, advocates of this position doubt the validity or originality of much of this recent research. Several criticisms of the monopolistic modelling logic underpinning 'new economic geography' have come from economic geography schools of thought as well as both orthodox and heterodox schools of economics. These critiques focus variously on the immeasurability of some of the notions of increasing returns inherent in the new economic geography frameworks, the static nature of some of its assumptions, the specific focus on the representative firm, the presence only of pecuniary economies and the absence of either human capital or technological spillovers as externalities. On the other hand, advocates of the new economic approaches argue that their analyses do provide insights into spatial economic phenomena which were previously unattainable under the existing analytical frameworks and toolkits. In this chapter we reflect on these current and future developments, putting equal weight on both evolutionary and institutional economic geographical conceptualizations as on geographical economic ones, set off against a historical review of agglomeration and regional economic growth theories. We show that the modern concepts and modern treatment of agglomeration and regional growth do indeed build upon previous classical insights, while at the same time also introducing new insights. This fusion of new and old ideas also raises new questions, poses new challenges, and opens up new directions for future research.

1.2 Classical and neoclassical insights into regional growth

The major developments in spatial economics and economic geography from the late nineteenth century up until the 1960s came from a variety of different traditions, and from a variety of different analysts. In terms of the location of economic activities, major insights were provided by amongst others Weber (1909), Lösch (1954 [1939]), Isard (1956) and Christaller (1933 [1966]). At the same time, related work on the causes and regional growth consequences of the spatial clustering of economic activities was also being undertaken by Lichtenberg (1960), and Vernon (1960) and Chinitz (1961) whose work focused specifically on issues relating to growth and agglomeration. In particular, the focus of their work was on the features of different types of agglomeration economies, and their analyses were undertaken within the traditional analytical framework of agglomeration phenomena, which had emerged as a fusion of the insights of Marshall (1890) and Hoover (1948). Marshall (1890) focused on the role of local knowledge spillovers, and the existence of non-traded local inputs and a local specialist labour pool, while Hoover

(1948), Ohlin (1933) and Isard (1956) allocated the sources of agglomeration advantages into internal economies of scale and external economies of scale in the form of localization and urbanization economies. Internal increasing returns to scale may occur to a single firm due to production cost efficiencies realized by serving large markets, and as such, there is nothing inherently spatial in this concept other than that the existence of a single large firm in space implies a large local concentration of factor employment. On the other hand, external economies are qualitatively very different.

Whether due to firm size or a large initial number of local firms, a high level of local factor employment may allow the development of external economies within the group of local firms in a sector. These are termed localization economies. The strength of these local externalities is assumed to vary, so that these are stronger in some sectors and weaker in others (Duranton and Puga, 2000). The associated economies of scale comprise factors that reduce the average cost of producing outputs in that locality. The theories on localization economies can be further enhanced by explicitly taking market form into consideration (Gordon and McCann, 2000). Externalities characterized by knowledge spillovers between firms in a spatially concentrated industry are generally known as Marshall–Arrow–Romer (MAR) externalities. The MAR theory in a dynamic context (Glaeser et al., 1992; Henderson et al., 1995) predicts, as Schumpeter (1934) did, that local monopoly is better for growth than local competition, because local monopoly restricts the flow of ideas to others and so allows innovator-internalization. Porter (1990) agrees with the importance of localization economies, also arguing that knowledge spillovers in specialized, geographically concentrated industries stimulate growth. On the other hand, urbanization economies reflect external economies passed to enterprises as a result of savings from the large-scale operation of the agglomeration or city as a whole, and which are therefore independent from industry structure. Relatively more populous localities, or places more easily accessible to metropolitan areas, are also more likely to house universities, industry research laboratories, trade associations and other knowledge-generating institutions. It is the dense presence of these institutions, which are not solely economic in character, but are social, political and cultural in nature, that support the production and absorption of know-how, stimulating innovative behaviour and differential rates of interregional growth (Harrison et al., 1997). The diverse industry mix in an urbanized locality therefore improves the opportunities to interact, copy and modify practices and innovative behaviour in the same or related industries. In her well-known theory on urban growth, Jane Jacobs (1969) defines diversity as a key source of agglomeration economies, and unlike the MAR theory, believes that the most important knowledge transfers come from outside the own industry.

Quigley (1998) describes four features of agglomeration economies. The first factor he describes concerns scale economies or indivisibilities within a firm, which are the historical rationale for the existence of productivity growth in agglomerated industries in the first place (Brakman et al., 2001; Isard, 1956). Without the existence of scale economies in production, economic activities would be dispersed so as to save transportation costs (Fujita and Thisse, 2002; Palivos and Wang, 1996). In consumption terms, the existence of public goods leads to urban amenities. Cities function as ideal institutions for the development of social contacts corresponding to various kinds of social and cultural externalities (Florida, 2002).

The second factor, namely shared inputs in production and consumption, encompasses the economies of localized industry described by Marshall. The use of shared inputs to

produce more differentiated consumption goods in agglomerations associated with variety, fashion, culture and style, is well known (Katz and Shapiro, 1985).

A third possible reason why agglomeration economies may provide greater economic efficiency growth arises from potential reductions in transaction costs (Martin and Ottaviano, 1999). The Western economies in general have developed primarily into services-based economies. Business and consumer services now make up most of urban employment nowadays, and most of these urban activities are characterized in terms of a knowledge-based information society. A logical outcome of the interaction between urban economies and knowledge-based service industries is the growing importance of transactions-based explanations of local economic productivity growth (Castells, 1989; Gottmann, 1983). The so-called Californian School of economic geography emphasizes transactional costs in explaining agglomeration economies (Scott, 1988), and the survival of local firms and the lower search costs of workers (Helsey and Strange, 1990; Kim, 1987; Acemoglu, 1996) demonstrate that in a matching-context between workers and firms returns to human capital accumulation can be shown to exist, even when all output in a city is produced with constant returns to scale and with no technological externalities. Again analogous to production, better matching may occur in consumer functions (shopping).

The fourth set of potential economies identified by Quigley (1998) relates to the application of the law of large numbers to the possibility of fluctuations in the economy. Fluctuations in purchases of inputs are usually as imperfectly correlated across firms, as the sales of outputs are across buyers. As such, less inventory holding is required due to the greater possibilities for the pooling of supplies.

Each of these aspects of agglomeration economies provides a possible rationale as to why regions characterized by agglomeration will generally exhibit higher growth than regions without such features. In addition to these features of agglomeration economies, there are also two additional features of cities which contribute to the growth potential of a city-region. Firstly, the structure of a regional or urban economy can be considered in a manner analogous to corporate diversification in product portfolios. Regional variety can be considered as a portfolio strategy to protect regional income from sudden asymmetric sector-specific shocks in demand (Mills, 1972; Attaran, 1986; Dissart, 2003). This will especially protect labour markets, and thus prevent sticky unemployment occurring. Even if interregional labour mobility is high, asymmetric shocks reduce economic growth as agglomeration economies and the tax base deteriorate (Krugman, 1993). Following this reasoning, industrial variety at the regional level would reduce regional unemployment and would promote regional economic growth, while specialization would increase the risk of unemployment and a growth slowdown. As for firms, a central question is whether related or unrelated diversification is most rewarding for stability and growth (Baldwin and Brown, 2004). One can expect that related industries more often (though, again, not as a rule) have correlated demand shocks. Therefore, spreading risk over unrelated sectors is likely to be preferred from the viewpoint of a portfolio strategy. However, one should take into account the possible benefits from related diversification as well. Analogous to economies of scope at the firm level, one expects knowledge spillovers within the region to occur primarily among related sectors, and only to a limited extent among unrelated sectors. In terms of agglomeration theory, Jacobs externalities are expected to be higher in regions with a related variety of sectors than in regions with an unrelated variety of sectors (Frenken et al., 2007).

Secondly, as we will see shortly, technological development and the diffusion of knowledge and innovation are regarded as central to the modern concept of regional growth. However, the concept of knowledge diffusion across space in the economic geographical literature dates back some sixty years, beginning with the growth pole theory of Perroux (1950) which was subsequently embedded in geographical space by Boudeville (1966). Its main assumption is that economic growth, manifested in the form of innovations, is spread throughout a growth centre's hinterland to lower-order cities and localities nearby. Innovations and knowledge once generated in a certain central location are expected to spread among regions from one locality to its neighbours (Richardson, 1978; Parr, 1999). Hirschman (1958) distinguished two types of spillover effects associated with growth pole theory: backward linkages and forward linkages. The former effects are associated with activities that provide inputs to economic activities, drawing them towards the location where the clients are. The latter concern activities that use outputs by new activities or expanding existing activities that draw them towards locations where these existing activities are already (over-)represented. This can turn into backwash effects that are usually unanticipated, occurring when the growth pole attracts so much attention and cumulative growth that it drains the surrounding areas. Migration of workers towards the pole and the concentration of investment capital in the initial centre of innovation initiate the emergence of high-level urban services in the growth pole. This can then lead to a further polarization of economic growth, restricting growth elsewhere (Richardson, 1978). The existence of spread effects is based on the belief that the ongoing growth of the core location (the growth pole) will eventually lead to diseconomies of scale due to congestion and the appreciation of factor costs. A parallel stream of work also emerged from Vernon (1960) and Chinitz (1961) in which the role of cities as incubators of new firms and new ideas was regarded as critical. More recently, this theoretical framework has been applied in agglomeration studies of Henderson (1997) and Rosenthal and Strange (2001) on innovation intensity, and Henderson et al. (1995) and Van Oort and Atzema (2004) on employment growth. All these papers argue that there is an urban product cycle notion in that new products are more easily developed in large diverse metro areas with a diversified industrial structure and skill base, and particularly those with many corporate headquarters (Pred, 1977), whereas mature products eventually are decentralized to hinterland or peripheral areas.

After the period of rapid analytical developments up to the late 1960s associated with the quantitative revolution in economic geography and the microeconomic-based breakthroughs in regional science (Isard, 1956), outside of the specialist research field, widespread interest in spatial economic issues to a large extent waned in both economics and geography for a period of two decades. As such, it was another twenty years until a major resurgence of interest in spatial and regional economic issues was witnessed. This resurgence of interest was associated with the work of Paul Krugman (1991) and Michael Porter (1990), and both of these commentators not only borrowed from the existing insights, but also added new insights to their analyses.

1.3 The 1990s revolution: new economic geography and new growth theory

Prior to the development of new trade theory, traditional international trade theory was largely unable to explain either intra-industry, intranational or intra-regional trade. At the same time, gravity models suggested that most trade tended to be localized. The

development of new trade theory based on the Dixit–Stiglitz (1977) modelling framework subsequently led to renewed interest in both localized and intra-industry trade. These developments in international trade theory in turn led to a renewed modelling interest in spatial economics in the form of new economic geography, and regional economics as a whole subsequently experienced a resurgence via a combination of the developments in both new economic geography and also new growth theories.

New economic geography is based on the insights and analytical approaches that are common to new growth theory and new trade theory. As both new growth theory and new trade theory pre-date new economic geography, it is worthwhile to recap the basic features and insights of new economic geography's two antecedent literatures. In both of these strands of literature the dominant analytical approach is the modelling of imperfect competition and increasing returns to scale within the monopolistic competition framework of Dixit and Stiglitz (1977), in which utility is a function of variety. New trade theories now allowed for the modelling of inter- as well as intra-industry trade flows within a general equilibrium framework in which the structure of demand and supply is endogenously determined.

Krugman (1991) first applied this modelling framework to the question of geography under conditions of economies of scale and labour mobility, and reinterpreted Marshall's principle of externalities as stemming from the benefits of the pooling of the local labour supply and the demand for specialized non-tradable inputs. In these models, spatial concentration and dispersion were seen to emerge as a natural consequence of market interactions involving economies of scale at the level of the individual firm, with many of the results generated by these models being reminiscent of the results of central place theory and the rank–size rule (Fujita et al., 1999). Indeed, the cumulative causation characteristics of these models is in many ways akin to the processes described amongst others by Pred (1977) and in this respect the Krugman–Fujita–Venables work builds on most of the standard location theory (Dymski, 1996; Krugman, 1993).

This spatial version of the Dixit–Stiglitz monopolistic competition theory has since become a crucial element in many spatial economists' models on the location of economic activities (Abdel-Rahman, 1988; Fujita et al., 1999) and several key insights have emerged from this literature. Firstly, if internal economies of scale are strong and transportation costs are low, this induces a circularity that tends to keep geographic concentration in existence once established (compare Pred, 1977 and Myrdal, 1957 on their notions on cumulative causation). The reason is that manufacturers in the larger economic agglomerations have an advantage, since the size of local demand allows them to profit more from internal economies of scale, and hence they can afford higher nominal wages. A higher local demand for goods induces a greater range of variety of goods, which induce real income effects that attract new workers, consumers and firms. These developments are manifested in a greater range of local forward linkages (the supply of a greater variety of goods increases the worker's real income) and local backward linkages (a greater number of consumers attracts more firms) as pecuniary externalities create scale economies at the individual firm level that are transformed in increasing returns at the level of a location as a whole (Gianmarco et al., 2001). In general, this effect will be stronger as local demand is greater and internal economies of scale are higher.

Meanwhile, this observation of spatial industrial concentration is also consistent with the observation that some producers survive in peripheral locations. One reason is that

peripheral producers exhibit local advantages outside the large agglomeration due to higher transportation costs, which means that they face less competition for their local demand. A second reason is that negative externalities such as congestion and high land rents in the larger agglomerations (Quigley, 1998) may eventually lead to decreasing returns to scale in cities (Glaeser et al., 1995; Moomaw, 1985). If the industrial sector itself constitutes a principal source of demand for industrial products, and if transportation costs increase with distance, then firms will cluster because they produce under increasing returns. The existence of sufficiently high transportation costs therefore ensures that multiple clusters will exist instead of one monocentric city. As such, the pull of Krugman's pecuniary externalities balances the push of transportation costs. The ultimate equilibrium depends on the initial point of departure and the extent of economies of scale, and the level and structure of transportation costs (McCann, 2005). Equilibrium no longer automatically means that spatial units of observation converge in terms of regional growth (Kubo, 1995).

A second and related recent body of literature related to geography and space has been developed on the basis of the new or endogenous growth theories. These theories themselves are built on similar foundations to new trade theory and new economic geography (Barro and Sala-i-Martin, 1995), although they are different in that they do not treat time in a comparative static manner, but take growth over time and its determinants as the principal subjects of the analysis. According to this view, when individuals or firms accumulate new capital, they inadvertently contribute to the productivity of capital held by others. Such spillovers may occur in the course of investment in physical capital or human capital (Lucas, 1988). As Romer (1986, 1990, 1994) demonstrated, if the spillovers are strong enough, the private marginal product of physical or human capital can remain permanently above the discount rate, even if individual investments would face diminishing returns in the absence of external boosts to productivity. These model approaches also became widely known as 'endogenous growth' theory, because technological change is also seen to be endogenously determined in these models (Romer, 1994; Solow, 1994).

When applied to regions and geography, these models all assume that the notion of increasing returns is spatially embodied in agglomeration economies. Endogenous regional growth models are similar to new economic geography models in that such effects can only operate within an environment of imperfectly competitive monopolistic competition. However, these regional growth models are also different to mainstream new economic geography models in that in the endogenous growth framework, local external economies may not only be associated with market size or pecuniary external economies, but can also be related to information or technological externalities and spillovers (Englmann and Walz, 1995; Rutten and Boekema, 2007). Martin and Ottaviano (1996) and Baldwin and Forslid (1997) show that by incorporating research and development (R&D) activity into models reminiscent of Krugman (1991) and Krugman and Venables (1996), local factor accumulation can play a similar role to that of either labour migration (Krugman, 1991) or input–output linkages (Puga and Venables, 1996; Venables, 1999) fostering agglomeration via local demand linkages. However, whereas agglomeration in new trade theory and new economic geography is the geographic outcome of modelling, in new growth theory it forms an endogenously determined explanation of growth. These types of arguments therefore provide some additional possible explanations for systematic variations in competitive advantage (Porter, 1998) across regions and why it is that

certain regions are able to maintain and even reinforce their advantages over other regions, once certain locations have taken a lead in a particular activity (Arthur, 1994; Krugman, 1991).

1.4 Economic geography and evolutionary economics

Several criticisms of the monopolistic modelling logic underpinning new economic geography have come from economic geography schools of thought (Martin and Sunley, 1996; Martin, 1999) as well as both orthodox (Neary, 2001) and heterodox schools of economics (Peneder, 2001). These critiques focus variously on the immeasurability of some of the notions of increasing returns inherent in these frameworks, the static nature of some of the assumptions, the specific focus on the representative firm, the presence only of pecuniary economies and the absence of either human capital or technological spillovers as externalities, and the problems associated with the iceberg transport costs assumption (McCann, 2005; Fingleton and McCann, 2007). Other evolutionary critiques (Martin and Sunley, 2003) also question the originality and validity of the Porter (1990) concept of clusters. It is fair to say, however, that many of these criticisms actually relate to specific models and specific papers, rather than to the whole field. On the other hand, the most fundamental critique of these fields in general relates to the question of institutions, and the relationship between knowledge and institutions. Within economics, institutions are regarded as being important in explaining economic growth (North, 1990; Aghion and Howitt, 1998; Helpman, 2004). However, for economic geographers and heterodox economists working within the evolutionary and institutional economics arenas, the role played by institutions in economic development is seen to be paramount. In this evolutionary–institutional schema, regions and countries that have more efficient institutions are therefore superior in both the generation and the diffusion of knowledge, and consequently have better prospects for economic growth. As such, while new economic geography and new growth theories are mathematically complex, they are still regarded by these analysts as being philosophically too simplistic. This is because they aim to produce generalizable predictions based on a representative model, whereas the counter-argument implies that the appropriate investments, favourable institutional arrangements and entrepreneurial dynamics which allow regions to grow are features of regions which have emerged for historically contingent and spatially contingent reasons, rather than generalizable reasons. For economic geographers, as well as institutional and evolutionary economists working in this tradition, cultural and cognitive proximity are therefore deemed to be just as important as geographical proximity in the transmission of ideas and knowledge (Boschma, 2005). Boschma and Lambooy (1999) further argue that the generation of local externalities is also crucially linked to the importance of selection in terms of 'fitness' of a local milieu, the sociological dimensions of which can be institutional, cultural, legal and historical. According to these perspectives, it is these specific historically contingent and geographically contingent features, rather than simply space as a dimension, which are crucial in determining the geography of entrepreneurship and growth (Audretsch et al., 2006).

The original behavioural geographical literature (Pred, 1966; Webber, 1964) focused on incomplete information, the limited cognitive capacities of entrepreneurs and the differences in information absorption abilities of firms at different stages in their life cycles (Alchian, 1950). However, institutional structures are now regarded as being much more

than simply the aggregation of individual choices, but rather the result of many interactive processes. Economic geography research has always emphasized the untraded interdependencies (Storper, 1997) that function as externalities and spillovers, and this has led to calls for research to focus on institutional issues (Amin and Thrift, 2002). As such, evolutionary economic geography theory focuses primarily on the creation of new spatial structures, rather than on explaining equilibrium states. Within the same spatial and institutional context, firms and entrepreneurs may arrive at different location behaviour either by means of chance occurrences or by fundamental processes of neo-Schumpeterian, creative destruction. Alternatively, different spatial and institutional contexts will mean that firms and entrepreneurs may arrive at either different or similar locational outcomes, but for a variety of different reasons. As such, the initial states which determine allocations may vary significantly, although the future trajectories of these initial outcomes are determined primarily by path-dependency phenomena, which themselves are underpinned by local externalities and spillovers. In turn, these path-dependent phenomena subsequently give rise to localized regional clustering.

Evolutionary economic theory, as originally developed by Nelson and Winter (1983), emerged from economics as a result of dissatisfaction with many of the equilibrating notions of neoclassical economics. In many ways these evolutionary theories are inspired by Darwinian processes of biological change (Boschma and Lambooy, 1999) and, as such, embody within themselves a very particular set of behavioural and environmental heuristics. Firstly, uncertainty provokes firms into routinized, risk-averse behaviour which determines to a large extent the available options and probable outcomes of searches. This implies that technical, technological and human capital issues generally exhibit path-dependency behaviour characterized by lock-in processes (David, 1985). Under these conditions, Arthur (1994) shows that the notion of increasing returns provides an explanation for why technology is able to maintain and reinforce its competitive advantage once it has taken the lead in the market, irrespective of whether the lead was taken due to superiority, coincidence or luck. Secondly, physical capital investments are a source of locational inertia. History, in the form of sunk costs resulting from the operation of many firms at a site, creates a first-mover disadvantage that can prevent relocation (Arthur, 1989; Rauch, 1993). Thirdly, the selection environment functions as a filtering mechanism that ultimately decides which of the innovations will thrive or fail. This selection environment consists of both a number of consumer and financial markets as well as a set of non-market institutions such as regulations, values, norms and customs (Storper, 1997). Evolutionary theory therefore implies that there may be a multiplicity of future spatial outcomes, many of which cannot be hypothesized on the basis of current observations. This argument is actually reflected in some of the new economic geography-type frameworks employed by several authors (Rauch, 1993; Bostic et al., 1997; Ottaviano and Puga, 1998; Berliant and Konishi, 2000; Cronon, 1991), who also conclude that there is a strong tendency toward path-dependency based on historical contingency. However, because of the long-lasting geographical history of cumulative causation since the 1950s, economic geographers do claim that this type of thinking and this approach to the treatment of history is theirs, in contrast to the 'newly' discovered nature of this subject in the mainstream economics literature.

The present weakness, however, with evolutionary and institutional approaches to regional growth, is that it is empirically primarily an *ex post* analytical framework. The

reason is that as yet, it is currently very difficult, if not impossible, to determine which observable outcomes can be more widely generalized or predicted on the basis of *ex ante* observations. In this sense, while many aspects of the growth processes can be described in detail, the ability to extrapolate is currently very limited.

1.5 Common ground?

Although the differences between the formal modelling approaches of new economic geography, new growth theory and the evolutionary–institutional approaches to regional growth at first may appear to be irreconcilable, common ground between these different competing theories can be found on several key points. Firstly, in each of these different literatures, as we have already seen, the role of agglomerations is regarded as being a crucial element of regional performance, and the common element here is the issue of local knowledge generation, accumulation and spillovers. Secondly, and related to the first point, is the issue of the level of connectivity, and specifically, the number of connections between local regional nodes to other key international nodal points in the global economy is regarded by all of these theories as being important (Saviotti, 1996). Recent work on global cities (Sassen, 2001, 2002; Taylor, 2004) suggests that particular cities that are well connected via international hub airports in particular, are nowadays consistently at an advantage over other locations in terms of acquiring relevant knowledge spillovers. Thirdly, the geographical scale over which knowledge spillovers operate is regarded as a critical issue, and once again, most of the apparently competing theories are largely in agreement.

On this third point, one of the features which neither the new economic geography nor the new growth theory explicitly models is the actual geographical scale over which any knowledge spillover mechanisms operate. As Jaffe et al. (1993) conclude, we know very little about where such spillovers actually go, although we can acquire some information regarding this point by studying the geographic location of patent citations. Jaffe et al. (1993) therefore test the extent to which knowledge spillovers are geographically localized. Their measured effects were particularly significant at the local Standard Metropolitan Statistical Areas (SMSA) level, indicating that localization fades over time, but only very slowly. Further research by Audretsch and Feldman (1996), Acs (2002) and Feldman (1994), amongst others, provides corroborating evidence that knowledge spillovers tend to be geographically bounded within the location where the new economic knowledge was created. Lucas (1993) emphasizes that the most natural context in which to understand the mechanics of dynamic knowledge externalities and economic growth is in metropolitan areas, where the compact nature of the geographic unit facilitates communication and human capital accumulation. He argues that the only compelling reason for the existence of cities would be the presence of increasing returns to agglomerations of resources that make these locations more productive. This view of human capital as social input that induces productivity gains in cities has been further explored by others (Bostic et al. 1997; Henderson, 1986; Rosenthal and Strange, 2004; Cheshire and Duranton, 2004) who all argue that the microeconomic foundation of the external effect of human capital is the sharing of knowledge and skills between workers that occurs through both formal and informal interactions. The distinction between tacit and implicit knowledge bases as against explicit knowledge bases is deemed to be crucial here in terms of the ways that knowledge externalities are embodied in growth (implicit) and innovation (explicit)

externalities. Intuitively it seems clear that the higher the average level of human capital (knowledge) or the more spatially concentrated are the numbers of agents, the more 'luck' these agents will have with their meetings and the more rapid will be the diffusion and growth of knowledge (Rauch, 1993, p. 381). Storper and Venables (2005) used the concept of 'buzz' to denote that much communication between decision-makers is actually accidental and happens in various non-organized meetings. Other authors also emphasize the importance of accidental meetings (Fu, 2007; Charlot and Duranton, 2004), whereby complex information transmission via face-to-face contacts plays a crucial role, in addition to the provision of specialized services and labour supply. These features are argued to be dominant in cities (Duranton, 1999; Feser, 2002). This all therefore points to metropolitan areas as being the major locations where the productivity-enhancing effects of human capital primarily operate (Gaspar and Glaeser, 1998; Glaeser, 1999), and this pure agglomeration argument (Gordon and McCann, 2000) provides a natural explanation for higher wages as well as higher land rents in cities.

These observations, which emphasize the role played by the city as a knowledge and information environment, also largely accord with many of the explanations employed by the economic geography, institutional and evolutionary approaches. The original behavioural arguments generally pointed to large urban agglomerations as being superior incubator locations (Chinitz, 1961) to other places. This thinking has also heavily influenced contemporary economic geography thinking. The difference, however, is in terms of the emphases. The evolutionary–institutional approaches stress institutions and policy-makers (Amin and Thrift, 2002) on the assumption that in each observed case, the actual outcome of these externalities on productivity remains heavily dependent on the historical economic context (Bostic et al., 1997), the industrial structure (Moomaw, 1988; Glaeser et al., 1992) and the specific role played by face-to-face contact in local production processes (McCann, 2007). Therefore, when behavioural and evolutionary explanations for interregional economic development are taken seriously, primary attention is paid to the behavioural and entrepreneurial causes of agglomeration. The concept of externalities in this schema is therefore also related to the nature of information transmission mechanisms between actors in firms and the cognitive and interactive characteristics that determine the construction of locational preferences.

1.6 Conclusions

The new economic geography and new growth approaches rightly argue that their analyses do provide insights into spatial and economic phenomena which were previously unattainable under the existing analytical frameworks and toolkits. The conceptualizations of endogenous growth, monopolistic competition and increasing returns to scale triggered a new phase of development in economic modelling. By accepting that in reality spatial and firm-level heterogeneity are much greater than these present (still) general equilibrium models allow for, the phenomena of path-dependency and heterogeneous sectoral development trajectories emphasized by evolutionary economic arguments do not necessarily contradict the analytical outcomes of new economic geography and new growth models. Rather more complicated, however, is the question of the role played by institutions, and the complexity here arises from the fact that whereas neoclassical economics employs a minimalist definition of institutions in terms of property rights and firms, institutional approaches variously allow for a whole array of social, legal, political, historical,

geographical and cultural phenomena to be characterized as institutions. As such, there is currently no agreed parsimonious definition of institutions on the part of institutional approaches. Therefore, while the analytical problems are themselves very complicated, the problems posed by these definitional issues are actually more problematic for the evolutionary–institutional approaches than they are for the new economic geography approaches. The reason is that the new economic geography and new growth theories assume that the dominant growth mechanisms are economic in nature and determined by pricing and allocation outcomes, whereas the evolutionary–institutional arguments assume that these other institutional phenomena are dominant. As such, until evolutionary–institutional approaches develop ways of clearly defining the nature, characteristics and behavioural features and outcomes of institutions, any actual real-world observations will always suffer from the inherent methodological problem of observational equivalence (McCann, 2007). Therefore, while there is already much common ground between these various approaches to regional growth, much interesting work remains to be done in order to reconcile fully these different analytical approaches.

References

Abdel-Rahman, H. (1988), 'Product differentiation, monopolistic competition and city size', *Journal of Urban Economics*, **18**, 69–86.

Acemoglu, D. (1996), 'A microfoundation for social increasing returns in human capital accumulation', *Quarterly Journal of Economics*, **111**, 779–804.

Acs, Z.J. (2002), *Innovation and the Growth of Cities*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Aghion, P. and P. Howitt (1998), *Endogenous Growth Theory*, Cambridge, MA: MIT Press.

Alchian, A.A. (1950), 'Uncertainty, evolution and economic theory', *Journal of Political Economy*, **58**, 211–21.

Amin, A. and N. Thrift (2002), *Cities: Reimagining the Urban*, Cambridge: Polity Press.

Arthur, B. (1989), 'Competing technologies, increasing returns and lock-in by historical events', *Economic Journal*, **99**, 116–31.

Arthur, B. (1994), *Increasing Returns and Path Dependence in the Economy*, Ann Arbor, MI: University of Michigan Press.

Attaran, M. (1986), 'Industrial diversity and economic performance in US areas', *Annals of Regional Science*, **20**, 44–54.

Audretsch, D.B. and M.P. Feldman (1996), 'R&D spillovers and the geography of innovation and production', *American Economic Review*, **86**, 630–40.

Audretsch, D.B., M.C. Keilbach and E.E. Lehrmann (2006), *Entrepreneurship and Economic Growth*, Oxford: University Press.

Baldwin, J.T. and W.M. Brown (2004), 'Regional manufacturing employment volatility in Canada: the effects of specialisation and trade', *Papers in Regional Science*, **83**, 519–41.

Baldwin, R.E. and R. Forslid (1997), 'The core–periphery model and endogenous growth', CEPR Working Paper no. 1749, London.

Barro, R.J. and X. Sala-i-Martin (1995), *Economic Growth*, Cambridge, MA: MIT Press.

Berliant, M. and H. Konishi (2000), 'The endogenous formation of a city: population agglomeration and market places in a location-specific production economy', *Regional Science and Urban Economics*, **30**, 289–324.

Boschma, R.A. (2005), 'Proximity and innovation: a critical assessment', *Regional Studies*, **39**, 61–74.

Boschma, R.A. and J.G. Lambooy (1999), 'Evolutionary economics and economic geography', *Journal of Evolutionary Economics*, **9**, 411–29.

Bostic, R.W., J.S. Gans and S. Stern (1997), 'Urban productivity and factor growth in the late nineteenth century', *Journal of Urban Economics*, **41**, 38–55.

Boudeville, J.R. (1966), *Problems of Regional Economic Planning*, Edinburgh: University Press.

Brakman, S., H. Garretsen and C. van Marrewijk (2001), *An Introduction to Geographical Economics*, Cambridge: Cambridge University Press.

Castells, M. (1989), *The Informational City: Information Technology, Economic Restructuring and the Urban–Regional Process*, Oxford: Blackwell.

Charlot, S. and G. Duranton (2004), 'Communication externalities in cities', *Journal of Urban Economics*, **56**, 581–613.

Cheshire, P.C. and G. Duranton (2004), *Recent Developments in Urban and Regional Economics*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.
Chinitz, B.J. (1961), 'Contrasts in agglomeration: New York and Pittsburgh', *American Economic Review*, **51**, 279–89.
Christaller, W. (1933), *Central Places in Southern Germany*, Englewood Cliffs, NJ: Prentice Hall; reprint and translation 1966.
Cronon, W. (1991), *Nature's Metropolis: Chicago and the Great West*, New York: Norton.
David, P. (1985), 'Clio and the economics of QWERTY', *American Economic Review*, **75**, 332–7.
Dissart, J.C. (2003), 'Regional economic diversity and regional economic stability: research results and agenda', *International Regional Science Review*, **26**, 423–46.
Dixit, A.K. and J.E. Stiglitz (1977), 'Monopolistic competition and optimum product diversity', *American Economic Review*, **67**, 297–308.
Duranton, G. (1999), 'Distance, land and proximity: economic analysis and the evolution of cities', *Environment and Planning A*, **31**, 2169–88.
Duranton, G. and D. Puga (2000), 'Diversity and specialisation in cities: why, where and when does it matter?', *Urban Studies*, **37**, 533–55.
Dymski, G.A. (1996), 'On Krugman's model of economic geography', *Geoforum*, **27**, 439–52.
Englmann, F.C. and U. Walz (1995), 'Industrial centers and regional growth in the presence of local inputs', *Journal of Regional Science*, **35**, 3–27.
Feldman, M.P. (1994), *The Geography of Innovation*, Boston, MA: Kluwer Academic Publishers.
Feser, E.J. (2002), 'Tracing the sources of local external economies', *Urban Studies*, **39**, 2485–2506.
Fingleton, B. and P. McCann (2007), 'Sinking the iceberg? On the treatment of transport costs in new economic geography', in B. Fingleton (ed.), *New Directions in Economic Geography*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, pp. 168–203.
Florida, R. (2002), *The Rise of the Creative Class*, New York: Basic Books.
Frenken, K., F.G. van Oort and T. Verburg (2007), 'Related variety, unrelated variety and regional economic growth', *Regional Studies*, **41**, 685–97.
Fu, S. (2007), 'Smart café cities: testing human capital externalities in the Boston metropolitan area', *Journal of Urban Economics*, **61**, 86–111.
Fujita, M. and J.F. Thisse (2002), *Economics of Agglomeration: Cities, Industrial Location, and Regional Growth*, Cambridge: Cambridge University Press.
Fujita, M., P. Krugman and A. Venables (1999), *The Spatial Economy: Cities, Regions and International Trade*, Cambridge, MA: MIT Press.
Gaspar, J. and E.L. Glaeser (1998), 'Information technology and the future of cities', *Journal of Urban Economics*, **43**, 136–56.
Gianmarco, I., P. Ottaviano and J.F. Thisse (2001), 'On economic geography in economic theory: increasing returns and pecuniary externalities', *Journal of Economic Geography*, **1**, 153–79.
Glaeser, E.L. (1999), 'Learning in cities', *Journal of Urban Economics*, **46**, 254–77.
Glaeser, E.L., J.A. Scheinkman and A. Schleifer (1995), 'Economic growth in a cross-section of cities', *Journal of Monetary Economics*, **36**, 117–43.
Glaeser, E.L., H.D. Kallal, J.A. Scheinkman and A. Schleifer (1992), 'Growth in cities', *Journal of Political Economy*, **100**, 1126–52.
Gordon, I.R. and P. McCann (2000), 'Industrial clusters: complexes, agglomeration and/or social networks?', *Urban Studies*, **37**, 513–32.
Gottmann, J. (1983), *The Coming of the Transactional City*, Maryland: University of Maryland and Institute for Urban Studies.
Harrison, B., M.R. Kelley and J. Gant (1997), 'Innovative firm behavior and local milieu: exploring the intersection of agglomeration, firm effects, and technological change', *Economic Geography*, **72**, 233–58.
Helpman, E. (2004), *The Mystery of Economic Growth*, Cambridge, MA: Harvard University Press.
Helsey, R.W. and W.C. Strange (1990), 'Matching and agglomeration economies in a system of cities', *Regional Science and Urban Economics*, **20**, 189–212.
Henderson, J.V. (1986), 'Efficiency of resource usage and city size', *Journal of Urban Economics*, **19**, 47–70.
Henderson, J.V. (1997), 'Externalities and industrial development', *Journal of Urban Economics*, **42**, 449–70.
Henderson, J.V., A. Kuncoro and M. Turner (1995), 'Industrial development in cities', *Journal of Political Economy*, **103**, 1067–85.
Hoover, E.M. (1948), *The Location of Economic Activity*, New York: McGraw-Hill.
Isard, W. (1956), *Location and Space-Economy: A General Theory Relating to Industrial Location, Market Areas, Land Use, Trade and Urban Structure*, Cambridge, MA: MIT Press.
Jacobs, J. (1969), *The Economy of Cities*, New York: Vintage.
Jaffe, A.B., M. Trajtenberg and R. Henderson (1993), 'Geographic localization of knowledge spillovers as evidenced by patent citations', *Quarterly Journal of Economics*, **36**, 577–98.

Katz, M.L. and C. Shapiro (1985), 'Network externalities, competition and compatibility', *American Economic Review*, **75**, 424–40.
Kim, S. (1987), 'Diversity in urban labor markets and agglomeration economies', *Papers of the Regional Science Association*, **62**, 57–70.
Krugman, P.R. (1991), 'Increasing returns and economic geography', *Journal of Political Economy*, **99**, 483–99.
Krugman, P. (1993), 'On the relationship between trade theory and location theory', *Review of International Economics*, **12**, 110–22.
Krugman, P. and A.J. Venables (1996), 'Integration, specialization and adjustment', *European Economic Review*, **40**, 959–67.
Kubo, Y. (1995), 'Scale economies, regional externalities and the possibility of uneven regional development', *Journal of Regional Science*, **35**, 29–42.
Lichtenberg, R.M. (1960), *One Tenth of a Nation*, Cambridge, MA: Harvard University Press.
Lösch, A. (1954), *The Economics of Location*, New Haven, CT: Yale University Press; originally published in German in 1939.
Lucas, R.E. (1988), 'On the mechanics of economic development', *Journal of Monetary Economics*, **22**, 3–42.
Lucas, R.E. (1993), 'Making a miracle', *Econometrica*, **61**, 251–72.
Marshall, A. (1890), *Principles of Economics*, New York: Prometheus Books.
Martin, P. and J.P. Ottaviano (1996), 'Growth and agglomeration', CEPR discussion paper 1529, London: Centre for Economic Policy Research.
Martin, P. and G.I.P. Ottaviano (1999), 'Growing locations: industry location in a model of endogenous growth', *European Economic Review*, **43**, 281–302.
Martin, R. (1999), 'The new "geographical turn" in economics: some critical reflections', *Cambridge Journal of Economics*, **23**, 65–91.
Martin, R. and P. Sunley (1996), 'Paul Krugman's geographical economics and its implications for regional development theory: a critical assessment', *Economic Geography*, **72**, 259–92.
Martin, R. and P. Sunley (2003), 'Deconstructing clusters: chaotic concept or policy panacea?', *Journal of Economic Geography*, **3**, 5–35.
McCann, P. (2005), 'Transport costs and new economic geography', *Journal of Economic Geography*, **5** (3), 305–18.
McCann, P. (2007), 'Sketching out a model of innovation, face-to-face interaction and economic geography', *Spatial Economic Analysis*, **2** (2), 117–34.
Mills, E.S. (1972), *Urban Economics*, Glenview, IL: Scott-Foresman & Co.
Moomaw, R.L. (1985), 'Firm location and city size: reduced productivity advantages as a factor in the decline of manufacturing in urban areas', *Journal of Urban Economics*, **17**, 73–89.
Moomaw, R.L. (1988), 'Agglomeration economies: localization or urbanization?' *Urban Studies*, **25**, 150–61.
Myrdal, G. (1957), *Economic Theory and Under-developed Regions*, London: Duckworth.
Neary, J.P. (2001), 'Of hype and hyperbolas: introducing the new economic geography', *Journal of Economic Literature*, **39**, 536–61.
Nelson, R.R. and S.G. Winter (1983), *An Evolutionary Theory of Economic Change*, Cambridge, MA: Belknap Press.
North, D.C. (1990), *Institutions, Institutional Change and Economic Performance*, Cambridge: Cambridge University Press.
Ohlin, B. (1933), *Interregional and International Trade*, Cambridge, MA: Harvard University Press.
Ottaviano, G.I.P. and D. Puga (1998), 'Agglomeration in the global economy: a survey of the new economic geography', *World Economy*, **21**, 707–31.
Palivos, T. and P. Wang (1996), 'Spatial agglomeration and endogenous growth', *Regional Science and Urban Economics*, **26**, 645–69.
Parr, J.B. (1999), 'Growth-pole strategies in regional economic planning: a retrospective view. Part 1: origins and advocacy', *Urban Studies*, **36**, 1195–216.
Peneder, M. (2001), *Entrepreneurial Competition and Industrial Location: Investigating the Structural Patterns and Intangible Sources of Competitive Performance*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.
Perroux, F. (1950), 'Economic space: theory and applications', *Quarterly Journal of Economics*, **64**, 89–104.
Porter, M. (1990), *The Competitive Advantage of Nations*, New York: Free Press.
Porter, M. (1998), *On Competition: Competing across Locations*, Cambridge, MA: Harvard Business School Press.
Pred, A.R. (1966), *The Spatial Dynamics of US Urban–Industrial Growth 1800–1914: Interspective and Theoretical Essays*, Cambridge, MA: The MIT Press.
Pred, A. (1977), *City-Systems in Advanced Economies: Past Growth, Present Processes and Future Development Options*, London: Hutchinson.
Puga, D. and A. Venables (1996), 'The spread of industry: spatial agglomeration in economic development', *Journal of the Japanese and International Economies*, **10**, 440–64.

Quigley, J.M. (1998), 'Urban diversity and economic growth', *Journal of Economic Perspectives*, **12**, 127–38.
Rauch, J.E. (1993), 'Does history matter only when it matters little? The case of city-industry location', *Quarterly Journal of Economics*, **20**, 843–67.
Richardson, H.W. (1978), *Regional and Urban Economics*, Hindsdale, IL: Dryden Press.
Romer, P.M. (1986), 'Increasing returns and long-run growth', *Journal of Political Economy*, **94**, 1002–37.
Romer, P.M. (1990), 'Endogenous technological change', *Journal of Political Economy*, **98**, S71–102.
Romer, P.M. (1994), 'The origins of endogenous growth', *Journal of Economic Perspectives*, **8**, 3–22.
Rosenthal, S.S. and W.C. Strange (2001), 'The determinants of agglomeration', *Journal of Urban Economics*, **59**, 191–229.
Rosenthal, S.S. and W.C. Strange (2004), 'Evidence on the nature and sources of agglomeration economics', in J.V. Henderson and J.F. Thisse (eds), *Handbook of Regional and Urban Economics: Cities and Geography*, Amsterdam: North Holland, pp. 2119–72.
Rutten, R. and F. Boekema (2007), *The Learning Region*, Cheltenham: Edward Elgar.
Sassen, S. (2001), *The Global City*, Princeton, NJ: Princeton University Press.
Sassen, S. (ed.) (2002), *Global Networks: Linked Cities*, London: Routledge.
Saviotti, P.P. (1996), *Technological Evolution, Variety and the Economy*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.
Schumpeter, J. (1934), *The Theory of Economic Development*, Cambridge, MA: Harvard University Press.
Scott, A.J. (1988), *New Industrial Spaces: Flexible Production Organization and Regional Development in North America and Western Europe*, London: Pion.
Solow, R.M. (1994), 'Perspectives on growth theory', *Journal of Economic Perspectives*, **8**, 45–54.
Storper, M. (1997), *The Regional World: Territorial Development in a Global Economy*, New York: Guildford Press.
Storper, M. and A.J. Venables (2005), 'Buzz: face-to-face contact and the urban economy', *Journal of Economic Geography*, **4**, 351–70.
Taylor, P.J. (2004), *World City Network: A Global Urban Analysis*, London: Routledge.
Van Oort, F.G. and O.A.L.C. Atzema (2004), 'On the conceptualization of agglomeration economies: the case of new firm formation in the Dutch ICT sector', *Annals of Regional Science*, **38**, 1–28.
Venables, A.J. (1996), 'Equilibrium locations of vertically linked industries', *International Economic Review*, **37**, 341–59.
Vernon, R. (1960), *Metropolis 1985*, Cambridge, MA: Harvard University Press.
Webber, M.M. (1964), 'The urban place and the nonplace urban realm', in M.M. Webber (ed.), *Explorations into Urban Structure*, Philadelphia, PA: University of Pennsylvania Press, pp. 79–153.
Weber, A. (1909), *Theory of the Location of Industries*, Chicago, IL: University of Chicago Press.

2 Space, growth and development

Roberta Capello

2.1 Economics and space

After 50 years of its existence, regional economics embraces a large number of approaches, theories and models for the interpretation of location choices and regional development trajectories. An increasing interpretative power characterizes the different models and theories once a historical perspective is assumed. The increasing interpretative capacity of the theoretical approaches can be attributed – among other factors – to the changes in the way space is inserted into the theoretical models. The aim of this chapter is to revisit – in a historical perspective – the different theoretical contributions, highlighting the evolution in the conceptualization of space, the different interpretations of growth so far provided by the different approaches, and the distinction between growth and development theories.

Economic activity arises, grows and develops in space. Firms, and economic actors in general, choose their locations in the same way as they choose their production factors and their technology. Productive resources are distributed unevenly in space: they are frequently concentrated in specific places (regions or cities) while they are entirely or partly non-existent in others. Quantitative and qualitative imbalances in the geographical distribution of resources and economic activities generate different factor remunerations, different levels of wealth and well-being, and different degrees of control over local development. The problem of factor allocation – which economists have conventionally treated as being the efficient allocation of the factors among various types of production – is more complex than this, in fact; and it is so because the spatial dimension is of crucial importance.

Space influences the way an economic system works. It is a source of economic advantages (or disadvantages) such as high (or low) endowments of production factors. It also generates geographical advantages, like the easy (or difficult) accessibility of an area, and a high (or low) endowment of raw materials. Space is also the source of advantages springing from the cumulative nature of productive processes in space: in particular, spatial proximity generates economies that reduce production costs (for example the transportation costs of activities operating in closely concentrated *filières*) and, in more modern terms, transaction costs (for example the costs of market transactions due to information gathering). These considerations highlight the need to supersede the purely allocative approach typical of a static interpretation of economic phenomena with a dynamic, indeed evolutionary, approach which ties allocative decisions to processes of development. The geographic distribution of resources and potential for development is only minimally determined by exogenous factors (raw materials, natural advantages). To a much larger extent, it results from past and recent historical factors: human capital, social fixed capital, the fertility of the land (due to the work of man) and accessibility (measured as the weighted distance from the main centres of production and consumption).

Regional economics is the branch of economics which incorporates the dimension 'space' into analysis of the working of the market. It does so by including space in logical schemes, laws and models which regulate and interpret the formation of prices, demand, productive capacity, levels of output and development, growth rates and the distribution of income in conditions of unequal regional endowments of resources. Furthermore, regional economics moves from 'space' to 'territory' as the main focus of analysis when local growth models include space as an economic resource and as an independent production factor, a generator of static and dynamic advantages for the firms situated within it – or, in other words, an element of fundamental importance in determining the competitiveness of a local production system.

It may seem somewhat trivial to emphasize the importance of space for economic activity. And yet, only recently has it been given due consideration by economic theory. Indeed, in the history of economics, analysts have devoted most of their attention and effort to determine the quantities of resources to be used for various purposes; they have concerned themselves with where those resources and activities are located or where they will be located only in the recent past. Analytical precedence and priority has thus been given to the temporal dimension over the spatial one.

There are several reasons for this belated consideration of space by economists. Firstly, as often pointed out by the founder himself of regional economics, Walter Isard,[1] the neoclassical school has conceived the temporal analysis of economic development as crucial and has always neglected the variable 'space' as a consequence – often in order to simplify the treatment. As Marshall wrote: 'The difficulties of the problem depend chiefly on variations in the area of space, and the period of time over which the market in question extends; the influence of time being more fundamental than that of space' (Marshall, 1920, Vol. 5, Chapter 15, section 1). Secondly, the treatment of the variable 'space' in economic analysis – especially if it is included in a dynamic approach – complicates the logical framework. The analytical tools until recently available to economists could not handle temporal and spatial dynamics simultaneously. Nor were they able to cope with the non-linearity of spatial phenomena like agglomeration or proximity economies. Finally, introduction of the variable 'space' required the discarding of the simplifying hypotheses (always dear to economists) of constant returns and perfect competition. According to the logic of a spatial market divided among producers, firms do not compete with all other firms, but only with those closest to them. Spatial distance is thus a barrier to entry which imposes a system of monopolistic competition – which also has only recently been formalized in analytical growth models.[2]

Two large groups of theories make up regional economics:

- location theory, the oldest branch of regional economics, first developed in the early 1900s, which deals with the economic mechanisms that distribute activities in space;
- regional growth (and development) theory, which focuses on spatial aspects of economic growth and the territorial distribution of income.

Location theory gives regional economics its scientific-disciplinary identity and constitutes its theoretical-methodological core. It typically has microeconomic foundations and it adopts a traditionally static approach. It deals with the location choices of firms and households. Linked with it are a variety of metaphors, cross-fertilizations and theoretical

inputs (from macroeconomics, interregional trade theory, development theory, mathematical ecology, systems theory) which have refined the tools of regional economics and extended its range of inquiry. In microeconomic terms, location theory involves investigation into the location choices of firms and households; but it also involves analysis of disparities in the spatial distribution of activities – inquiry which enables interpretation of territorial disequilibria and hierarchies. Location theory uses the concepts of externalities and agglomeration economies to shed light on such macro-territorial phenomena as disparities in the spatial distribution of activities, thereby laying the territorial bases for dynamic approaches.

Regional growth theory is instead intrinsically macroeconomic. However, it differs from the purely macroeconomic approaches of political economy in its concern with territorial features. Just as we speak of the micro-foundations of macroeconomics, so we may speak of the locational foundations of regional growth theory.

Numerous cross-fertilizations have taken place between these two branches of regional economics, and they have brought the traditional notions of space on each side – physical-metric for location theory, uniform-abstract for regional growth theory – closer together. The recent conception of space used in local development theories can be defined as diversified-relational: this is the bridge and the point of maximum cross-fertilization between the two traditional branches of regional economics. It yields an authentic theory of regional development based on the intrinsic relationalities present in local areas. These three conceptions of space are still today separate, however, and their integration has only been partly accomplished by the more modern notion of diversified-stylized space used by recent theories of local growth.

This chapter presents in detail the different notions of space in the different theories, through which a clear definition of the interpretative capacity of the theory emerges; physical-metric in section 2.2, uniform-abstract in section 2.3, diversified-relational in section 2.4, and diversified-stylized in section 2.5.[3] Moreover, within regional growth theories, very different conceptualizations of growth have been developed. The identification of the real meaning of growth brings about two main advantages; firstly, it prevents the attribution to theories and models of aims that they do not in fact set for themselves; secondly, the distinction drawn by the above classification of conceptions of growth dispels some apparent contradictions in theories and models of regional development (section 2.6). Finally, the new models of regional growth theories – rooted in complexity theory – embed non-linearities and cumulative self-reinforcing mechanisms, with the result that the univocity and mechanism of the results of the original neoclassical and Keynesian regional growth theories are abandoned. As a consequence, the distinction between regional divergence and convergence theories is by far superseded (section 2.7).

2.2 Location and physical-metric space

The first and earliest group of theories in regional economics falls under the heading of 'location theory'. This group adopts a purely geographical conception of continuous, physical-metric space definable in terms of physical distance and transportation costs. Thus interpreted are the regularities of price and cost variations in space, and their consequences in terms of location choices and the dividing of the market among firms. This was the conception of space used by the great geographers of the first half of the twentieth century.[4]

Location theory seeks to explain the distribution of activities in space, the aim being to identify the factors that influence the location of individual activities, the allocation of different portions of territory among different types of production, the dividing of a spatial market among producers, and the functional distribution of activities in space. These various phenomena are analysed by removing any geographical (physical) feature that might explain the territorial concentration of activities,[5] so that location choices are interpreted by considering only the great economic forces that drive location processes: transportation costs, which diffuse activities in space, and agglomeration economies, which instead cause activities to concentrate. By balancing these two opposing forces, these models are able to account for the existence of agglomerations of economic activities even on the hypothesis of perfectly uniform space.

Location models differ according to hypotheses on the spatial structure of demand and supply which reflect the aims that the models pursue. There are models whose aim is to interpret the location choices of firms, on the assumption of punctiform final and raw materials markets with given locations. Choice of location is determined in this case by an endeavour to minimize transportation costs between alternative locations and under the influence of agglomeration economies (theories of minimum-cost location). Here the obligatory reference is to the models developed by Alfred Weber and Melvin Greenhut. There are then models which seek to identify the market areas of firms, that is, the division of a spatial market among producers. In this case, the models hypothesize a demand evenly distributed across the territory which determines the location choices of firms, these being assumed to be punctiform. Locational equilibrium is determined by a logic of profit maximization whereby each producer controls its own market area (theories of profit-maximizing location); the reference here being to the market area models developed by, for example, August Lösch and Harold Hotelling.[6]

There are then models which seek to identify production areas. That is, they seek to identify the economic logic whereby a physical territory (land) is allocated among alternative types of production. In this case, the models are based on assumptions about the structure of demand and supply which are the reverse of those made by theories of market areas. The final market is punctiform in space (the town or city centre), while supply extends across the territory. Activities are organized spatially according to access to the final market, and locational equilibrium arises from a balancing between transportation costs on the one hand, and the costs of acquiring land for a central location on the other. The models developed by Johann Heinrich Von Thünen, William Alonso and the 'new urban economics' school express this logic.[7]

Finally, location theory analyses the economic and spatial mechanisms that regulate the size of territorial agglomerations, their functional specialization and their territorial distribution. These models put forward a more complex and general theory of location and the structure of the underlying economic relations able to account for the existence of diverse territorial agglomerations within a framework of general spatial equilibrium. The principal contributions to development of this theory have been made by Walter Christaller and August Lösch.[8]

2.3 Regional growth and uniform-abstract space

The second large group of theories pertaining to regional economics seek to explain why growth and economic development come about at local level. In this case regional

economics analyses the capacity of a subnational system – a region, a province, a city, an area with specific economic features – to develop economic activities, to attract them, and to generate the conditions for long-lasting development. Here, by 'regional economic development' is meant the ability of a local economic system to find, and constantly to recreate, a specific and appropriate role in the international division of labour through the efficient and creative use of the resources that it possesses. By emphasizing the more economic elements of this definition, regional development can be defined as the ability of a region to produce, with a (comparative or absolute) advantage, the goods and services demanded by the national and international economic system to which it belongs.

The first theories of regional growth were developed midway through the twentieth century. They used a conception of space – as uniform-abstract, no longer physical and continuous but abstract and discrete – entirely different from the physical-metric space of location theory. Geographic space was divided into 'regions', areas of limited physical-geographical size (largely matching administrative units) considered to be internally uniform and therefore synthesizable into a vector of aggregate characteristics of a social-economic-demographic nature: 'small countries' in the terminology of international trade but, unlike nations, characterized by marked external openness to the movement of production factors.[9]

The advantage of this conception of space is that it enables the use of macroeconomic models to interpret local growth phenomena. But although these models fit the above-mentioned features, they nevertheless, and it seems inexorably, require the analyst to exclude any mechanism of interregional agglomeration, to discard location theory, to ignore the advantages of local proximity, and instead to assume unequal endowments of resources and production factors, unequal demand conditions and interregional disparities in productive structures as the determinants of local development. Space is thus no more than the physical container of development and performs a purely passive role in economic growth, while some macroeconomic theories reduce regional development to the simple regional allocation of aggregate national development.

Theories which take this view of space are growth theories developed to explain the trend of a synthetic development indicator – income, for instance. Although this approach inevitably entails the loss of qualitative information, its undeniable advantage is that it makes modelling of the development path possible. These theories differ sharply in their conceptions of growth: there are those which conceive growth as a short-term increase in output and employment, and others which instead identify the growth path in a long-period increase in output associated with higher levels of individual well-being (high wages and per capita incomes, more favourable prices on the interregional market).

This conception of space has been adopted by the neoclassical regional growth theory, the export-base theory, and the interregional trade theory which developed from various branches of mainstream economics in the 1950s and 1960s:[10] macroeconomics, neoclassical economics, development economics and economics of international trade.

2.4 Local development and diversified-relational space

The definition of a diversified-relational space

Whilst the theories developed within a uniform-abstract space use the term 'space' to denote territorial areas assumed to be internally homogeneous and uniform, other

theories conceive 'space' as diversified. This change of perspective allows economic activities and production factors, demand and sectoral structure, to be treated as spatially heterogeneous within a region, so that territorial relations are cast in new light.

This new conception of space enables identification of highly distinct polarities in a territory. Activities, resources, economic and market relations structure themselves around these polarities to generate a cumulative process of territorial agglomeration and a virtuous circle of development. This conception of space restores one of the inspiring principles of location theories – that of agglomeration economies as the source of local development – to theories of regional development. It is evident that any connection with geographical space, abstract or administrative, is thus severed. A more complex conception of space takes over, one based on the economic and social relations that arise in a territorial area. Whence derives the expression 'diversified-relational space'.

When space is conceived as 'diversified-relational', theories radically change in their nature. A macroeconomic and macro-territorial approach gives way to a micro-territorial and micro-behavioural one. The notion of a region as a portion of a national system acting and reacting economically as a single, internally homogeneous system is abandoned. Its place is taken by individual economic actors (large or small, public or private, multinational or local) whose behaviour is studied in terms of location choices, productive and innovative capacity, competitiveness, and relations with the local system and the rest of the world.

The qualitative nature of theories – only in recent years superseded thanks to the more advanced and sophisticated modelling techniques at the basis of theories examined also in Chapters 4 and 5 of this volume[11] – led in the mid-1970s to the distinction in the literature between ' "pure and exact" regional theory without agglomeration economies, on the one hand, and "applied regional theory" which is inexact but takes agglomeration factors into account, on the other hand' drawn by Edwin Von Böventer.[12]

The theories within a diversified-relational space approach abandon the short-run view of development as a simple increase in income and employment, and also that of individual well-being, and assume a longer-term perspective. They identify all the tangible and intangible elements in a local area which determine its long-term competitiveness and enable it to maintain that competitiveness over time.

The theories analysed with this conception of space seek to identify the factors which render the costs and prices of production processes lower than they are elsewhere. These factors are: (1) elements exogenous to the local context, which originate externally to the area and are transferred into it either fortuitously or deliberately; and (2) endogenous elements which arise and develop within the area and enable it to initiate a process of self-propelling development.

Exogenous elements comprise the following: the fortuitous local presence of a dominant firm or a multinational company; the diffusion in the area of an innovation produced elsewhere; or the installation of new infrastructure decided by external authorities.[13] Although these elements have nothing to do with local features and productive capacities, once they are present in an area they may catalyse new economic activities and development. Endogenous elements are entrepreneurial ability and local resources for production (labour and capital); and in particular the decision-making capacity of local economic and social actors able to control the development process, support it during phases of transformation and innovation, and enrich it with external knowledge and information.

All these are factors strengthened and enhanced by a concentrated territorial organization which generates: local processes of knowledge-acquisition and learning; networks of economic and social relations which support more efficient and less costly transactions; and advantages of economic and physical proximity among economic actors.

The assumption of diversified space entails definitive abandonment of the notion that regional development consists solely in the allocation of resources among regions. Instead, regional development must be conceived as stemming from local productive capacity, competitiveness and innovativeness. The neoclassical model of interregional growth (Borts and Stein's one-sector model) presumed that the national growth rate is exogenously determined, and that the problem for regional development theory is explaining how the national growth rate is distributed among regions. According to this logic of competitive development, the growth of one region can only be to the detriment of the growth of another region, in a zero-sum game.[14] The theories examined here adopt a notion of generative development whereby the national growth rate is the sum of the growth rates achieved by individual regions. National economic development may well increase because of growth achieved by a particular territorial area, and this growth may also arise even in the presence of the same quantity of resources, thanks to increasing returns (as for the theories discussed in the next three chapters).

Interpretation of space as diversified-relational has restored to theories of regional development one of the key concepts of location theory – namely agglomeration economies – and made them the core of local development processes. According to this conception, which received its fullest development in the 1970s and 1980s, space generates economic advantages through large-scale mechanisms of synergy and cumulative feedback operating at local level.

A number of seminal theories of the early 1960s for the first time conceived space as diversified-relational. Development was defined, in the words of Perroux, as 'a selective, cumulative process which does not appear everywhere at the same time but becomes manifest at certain points in space with variable intensity'.[15] Perroux's definition affirmed the existence of 'poles' at which development concentrates because of synergic and cumulative forces generated by stable and enduring local input–output relations facilitated by physical proximity. Space is thus conceived as diversified and 'relational'.

But it was during the 1970s that studies on 'bottom-up' processes of development, on districts and local milieux, gave the notion of diversified-relational space its most thorough formulation. The conceptual leap consisted in interpreting space as 'territory', or in economic terms, as a system of localized technological externalities: a set of tangible and intangible factors which, because of proximity and reduced transaction costs, act upon the productivity and innovativeness of firms. Moreover, the territory is conceived as a system of local governance which unites a community, a set of private actors and a set of local institutions. Finally, the territory is a system of economic and social relations constituting the relational or social capital of a particular geographical space.[16]

Any connection with abstract or administrative space is thus obviously discounted. A more intangible account of space is adopted instead, which emphasizes – by focusing on the economic and social relations among actors in a territorial area – more complex phenomena which arise in local economic systems.

Precisely because the diversified-relational space theories of the 1970s and 1980s viewed development as depending decisively on territorial externalities in the form of location

and spatial proximity economies, they stressed (for the first time in the history of economic thought) the role of endogenous conditions and factors in local development. These theories adopted a micro-territorial and micro-behavioural approach; they can be called theories of development because their purpose was not to explain the aggregate growth rate of income and employment – as in the case of the above-mentioned uniform-abstract space theories – but instead to identify all the tangible and intangible elements of the growth process.

In the theories which conceived space as diversified-relational, location theory was inextricably and interestingly wedded with local development theory. By pointing out that concentration generates locational advantages, which in their turn create development and attract new firms whose presence further boosts the advantages of agglomeration, these theories elegantly revealed the genuinely 'spatial' nature of the development mechanism.

In this sense, diversified-relational space theories form the core of regional economics, the heart of a discipline where maximum cross-fertilization between location theory and development theory permits analysis of regional development as generative development – the national growth rate is the sum of the growth rates achieved by individual regions – as opposed to the competitive development envisaged by certain uniform-abstract space theories, where regional development is nothing but the simple regional allocation of aggregate national development.

The intriguing objective of these theories is to explain the competitiveness of territorial systems, the local determinants of development, and the capacity of an area to achieve and maintain a role in the international division of labour. They thus seek to identify the local conditions that enable an economic system to achieve and maintain high rates of development.

The active role of space on local development: agglomeration economies

Up to the 1970s, space was inserted into theories and models with two distinct roles: (1) the role of a physical barrier – or of a spatial friction – against economic activity, taking the form of the physical distance between input and output markets conceptualized by models as a generic transportation cost; (2) that of a 'physical container' of development, a simple geographical area often associated with the administrative region by aggregate macroeconomic theories – but also with smaller local areas (simple geographic agglomerations within a region, as envisaged by the more microeconomic theories examined in the previous chapter). In both cases, space plays no part in determining the development path of a local economy. The same economic logic explains the development of regions, metropolitan areas or, more generally, densely populated industrial areas. The export-base theory can be applied just as well to a region as to a country, with no change in the logic of its underlying reasoning. The Harrod–Domar model, too, and likewise the neoclassical growth models, fit both regional cases and national ones, which testifies to its aspatiality (Capello, 2007).

A radical change in the conceptualization of space which took place in the middle of the 1970s gives it a very different role in development. No longer a simple geographical container, space is conceived as an economic resource, as an independent production factor. It is the generator of static and dynamic advantages for firms, and a key determinant of a local production system's competitiveness. According to the theories examined

in this chapter, space is a source of increasing returns, and of positive externalities taking the form of agglomeration and localization economies. Higher growth rates are achieved by local production systems where increasing returns act upon local productive efficiency to reduce production and transaction costs, enhance the efficiency of the production factors and increase innovative capacity. Regional development consequently depends upon the efficiency of a concentrated territorial organization of production, not on the availability of economic resources or their more efficient spatial allocation.

This new conception of space has several implications. Space can only be a diversified space: in a diversified space it is possible to distinguish (even internally to a region) the uneven distribution of activities. If this is the case, development comes about selectively in areas where the concentrated organization of production exerts its positive effects on static and dynamic efficiency. At the same time, in this new conception space is relational, in that the economic and social relations which arise in an area perform crucial functions in various respects. They ensure the smoother operation of market mechanisms, more efficient and less costly production processes, the accumulation of knowledge in the local market and a more rapid pace of innovation – all of which are factors that foster local development.

On adopting this new notion of space it is no longer possible to treat development as exogenous in origin. Development is now by definition endogenous. It is fundamentally dependent on a concentrated organization of the territory, embedded in which is a socio-economic and cultural system whose components determine the success of the local economy: entrepreneurial ability, local production factors (labour and capital), relational skills of local actors generating cumulative knowledge-acquisition and, moreover, a decision-making capacity which enables local economic and social actors to guide the development process, support it when undergoing change and innovation, and enrich it with the external information and knowledge required to harness it to the general process of growth, and to the social, technological and cultural transformation of the world economy. The theories presented in this chapter accordingly endeavour to identify the genetic local conditions which determine the competitiveness of a local production system and ensure its persistence over time. They seek out the local factors which enable areas, and the firms located in them, to produce goods demanded internationally with an (absolute) competitive advantage, to maintain that advantage over time by innovating, and to attract new resources from outside.

Theories of local endogenous development divide into two broad strands. On the one hand neo-Marshallian inquiry, which views local growth as resulting from externalities acting upon the static efficiency of firms, has been expanding and consolidating for years. On the other, the neo-Schumpeterian literature, which has arisen more recently, interprets development as resulting from the impact of local externalities on the innovative capacity of firms.

The logical leap of interpreting space as an active factor in development forcefully imposed itself upon the history of economic thought in the early 1970s, when unprecedented patterns of local development in Italy surprised theoreticians by resisting explanation based on conventional models. During the early 1970s, the sudden and rapid growth achieved by certain Italian regions – those of the north-east and the centre in particular[17] – when the country's industrialized areas[18] were showing evident signs of economic crisis, could be explained neither by a neoclassical paradigm of interregional

mobility of production factors (which greatly decreased in those years), nor by a paradigm centred on large firm efficiency (à la Perroux), nor by a Keynesian paradigm of development driven by external demand.

Numerous neo-Marshallian theorists around the world pursued very similar lines of theoretical inquiry during the 1970s and 1980s (today there is still no lack of theory on the matter): Walter Stöhr developed the concept of 'bottom-up development', Enrico Ciciotti and Reinhart Wettmann that of 'indigenous potential', Bengt Johannison of 'local context', Bernardo Secchi and Gioacchino Garofoli of 'system areas', and Claude Courlet-Bernard Pecqueur and Bernard Ganne of 'localized industrial system'.[19] But the first systematic theory of endogenous development was produced in Italy by Giacomo Becattini with his seminal study on the 'Marshallian industrial district' published in the mid-1970s.[20] The theory of the industrial district – which originated in the work of the great neoclassical economist Alfred Marshall[21] – was the first to conceptualize external economies (of agglomeration) as sources of territorial competitiveness. It did so with a model in which the economic aspects of development are reinforced by a socio-cultural system which fuels increasing returns and self-reinforcing mechanisms of development.

These neo-Marshallian studies, in which space generates and develops mechanisms of productive efficiency, bred theories which identified the territory as the generator of dynamic external economies – that is, all those advantages which favour not only the productive efficiency of firms but also their innovative efficiency. In the neo-Schumpeterian strand of analysis on local development, space reduces the uncertainty associated with every innovative process.[22]

Finally, when space is viewed as generating advantages for firms, and therefore as an active component in the development process, scholars of local development shift their attention to the role of the urban space (the city) as the place where agglomeration economies are generated – be these localization or urbanization economies – and therefore as the place where the economic development of the entire region is rooted and structured. Hence, as the models of Christaller and Lösch show, the existence of an advanced and efficient city, and of an urban system organized into a network of vertical and horizontal relationships reflecting an efficient division of labour, may determine the success and development of a region.

2.5 Regional growth and diversified-stylized space: towards convergence?

Until the end of the 1980s the different conceptions of space – uniform-abstract and diversified-relational space – developed within regional economics without the slightest convergence between them.

The 1990s saw the development of more advanced mathematical tools for analysis of the qualitative behaviour of dynamic non-linear systems (bifurcation, catastrophe and chaos theory) together with the advent of formalized economic models which abandoned the hypotheses of constant returns and perfect competition. These advances made it possible to incorporate agglomeration economies – stylized in the form of increasing returns – into elegant models of a strictly macroeconomic nature.[23]

The reference is in particular to the models of 'new economic geography' and endogenous growth in which space becomes diversified-stylized.[24] These theories anchored their logic on the assumption that productive activities concentrate around particular 'poles' of development, so that the level and growth rate of income is diversified even within the

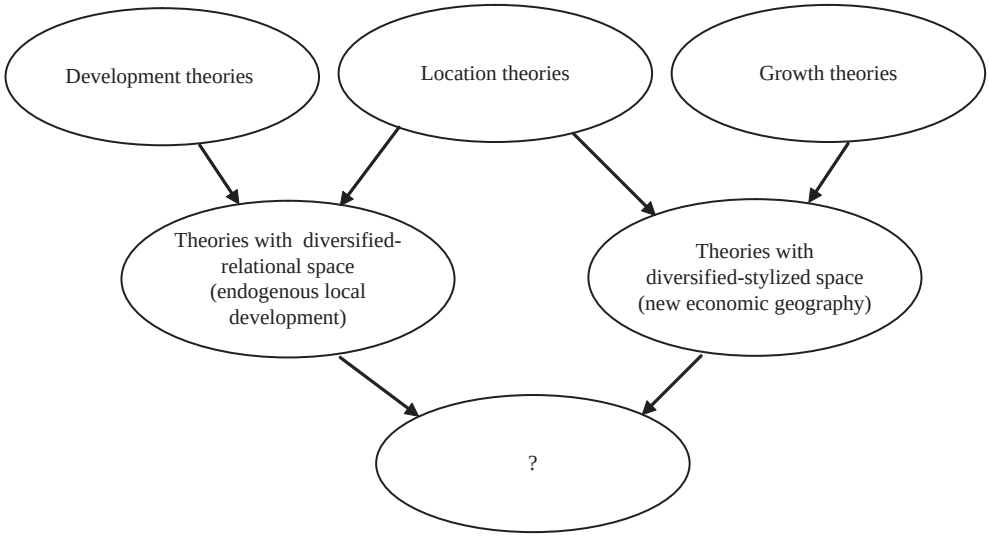
same region. Moreover, these models stylized areas as points or abstract dichotomies in which neither physical-geographical features (for example morphology, physical size) nor territorial ones (for example the local-level system of economic and social relations) play a role.

These theories achieved considerable success and acclaim in the academic community because they showed that territorial phenomena can be analysed using the traditional tools of economic theory (optimizing choices by individual firms and people), and that the various conceptions of space can – apparently – be synthesized. These models in fact conceived growth as an endogenous growth generated by the advantages of the spatial concentration of activities, and by the agglomeration economies typical of diversified space theories. They counterposed dynamic growth mechanisms with increasing returns and transportation costs, thus reprising the economic-locational processes analysed by location theory.

Though diversified (inasmuch as there exist territorial poles of concentrated development), space in these models is stylized into points devoid of any territorial dimension. Thus the notion of space as territory so favoured by regional economists is inevitably abandoned. This stylized space does not comprise localized technological externalities, nor the set of tangible and intangible factors which, thanks to proximity and reduced transaction costs, act upon the productivity and innovative capacity of firms; nor the system of economic and social relations constituting the relational or social capital of a particular geographical area. Yet these are all elements which differentiate among territorial entities on the basis of specifically localized features. As a consequence, these approaches are deprived of the most interesting, and in a certain sense intriguing, interpretation of space as an additional resource for development and as a free-standing production factor. Predominant instead is a straightforward, somewhat banal, view of space as simply the physical or geographical container of development.

This new conception of space has partly resolved the problem from which regional development theories have always suffered: their inability to construct formal models which combine specifically territorial features, like externalities and agglomeration economies, with macroeconomic laws and processes of growth. However, it should be pointed out that the assumption of a stylized rather than relational space deprives the polarities envisaged by such models of a territorial dimension able to give space – through synergy, cooperation, relationality and collective learning – an active role in the growth process. The introduction of agglomeration advantages in stylized form, through increasing returns, cancels out the territorial dimension. And in so doing it divests these theories of the aspect of greatest important to regional economists: namely space as territory defined as a system of localized technological externalities, or as a set of material and non-material factors which by virtue of proximity and reduced transaction costs act upon firms' productivity and innovativeness. Finding a way to incorporate the territorial dimension into theories already able to merge physical-metric, uniform-abstract and diversified space is the challenge that now faces regional economists.

To conclude, a certain convergence has come about between the large groups of theories discussed. Diversified-relational space theories, in particular those of (endogenous) local development, merge together ideas put forward by the theories of development and of location. Diversified-stylized space theories (in particular new economic geography) amalgamate growth and location theories (Figure 2.1). Nevertheless, still required is the



Source: Capello (2007).

Figure 2.1 Convergence among theoretical approaches

further step forward which would produce an approach combining the economic laws and mechanisms which explain growth, on the one hand, with the territorial features that spring from the intrinsic relationality present at local level on the other. Such an approach would represent the maximum of cross-fertilization among location theory, development theory and growth macroeconomics; a synthesis which would bring out the territorial micro-foundations of macroeconomic growth models (Figure 2.1). An undertaking of this kind, though, would require analysis of variables besides the cost of transport, which annuls the territory's role in the development process. Also necessary would be variables that give the territory prime place – even in purely economic models – among local growth mechanisms. This is the challenge that awaits regional economists in the years to come.

2.6 The different interpretations of regional growth and development

In the history of regional economics, no single definition has been given to the concept of regional growth. Rather, the various theories on the subject pertain to three 'philosophies' which have interpreted economic dynamics. The first, that of the classical (and neoclassical) economists of the eighteenth and nineteenth centuries, interprets the growth process in terms of productive efficiency, of the division of labour in a Smithian sense, and of production factor productivity, and hence examines the dynamics of wages, incomes and individual well-being. The second philosophy adopts a short-term view of growth and concentrates on the exploitation of given and unused capital resources and of large labour reserves. The third philosophy – the most modern of them – interprets the growth path as a problem concerning competitiveness and long-term dynamics and therefore takes the constant innovation of an economic system to be essential for development patterns.

We can use these three philosophies and their three views of the economic dynamic to classify the theories analysed later into three groups and highlight their normative aims:

1. The theories belonging to the first group aim to identify the factors that generate employment and income in a local system over the short term. They hypothesize the existence of unused production capacity (capital stock) and large labour reserves. In these conditions, local economic growth does not depend on the structure and dynamics of supply (which by definition is able to expand and respond rapidly to market requirements); rather, it is driven by growing demand for locally produced goods which exerts an income multiplier effect through increases in consumption and employment.[25] This was the definition given to growth by the first theories of the 1950s, which presupposed a problem of unemployment.
2. A second group of theories seeks to identify the economic mechanisms which enable a region to move out of poverty, start along a growth path, and ensure a certain level of well-being and per capita income for its inhabitants. Growth is a problem of individual well-being to be addressed in two ways: by acting upon factor productivity, thereby obtaining increases in real per capita wages and incomes; and by fostering processes of production specialization which yield advantages deriving from the purchase of goods on interregional markets at prices lower than they would be if the goods were produced internally to the region. These theories also comprise the notion of relative growth – of divergence/convergence in levels and rates of growth among regions – in that they measure the magnitude and trend of disparities among per capita incomes.[26] Growth was viewed in this way by most of the theories developed in the 1960s. Problems of poverty, underdevelopment and inequalities in the spatial distribution of income are the normative aspects of concern to these models.
3. The theories in the third group embrace a more modern conception of growth. They investigate the local conditions that enable the economic system to achieve high levels of competitiveness and innovativeness and, more crucially, to maintain those levels over time. Growth is defined as an increase in a region's real production capacity and its ability to maintain that increase. This conception is adopted by present-day theories and models of regional growth.

This classification is useful for two reasons. Firstly, it prevents the attribution to theories and models of aims that they do not in fact set for themselves. For example, it is wrong and misleading to think that theories which seek to identify processes of employment growth on the assumption of given but unused resources are able to suggest policies for long-term development. Indeed, it is hazardous to base normative action intended to foster a long-term dynamic on theories which concern themselves with the short period.

Secondly, the distinction drawn by the above classification of conceptions of growth dispels some apparent contradictions in theories and models of regional development. According to the conception of short-period income growth, an increase in exports is a development mechanism because it creates income. Yet from the viewpoint of individual well-being, it removes goods from final consumption and consequently hampers growth. Likewise, when development is viewed in terms of a short-term increase in income, emigration from a region is a cost because it deprives the area of effective demand (although it does so only at the level of subsistence consumption). But if the concern is with individual well-being, emigration is viewed as a positive factor in a region's development because it redresses imbalances (and consequently inefficiencies and income differentials) in the local labour market. On this view, surplus labour has nil marginal productivity and

tends to spend any increase in income on consumption, rather than on savings and production investments.[27] Far from being a resource for production development, it is an obstacle to growth, and its reduction statistically increases per capita income. Finally, if the focus is on an area's potential for long-period development, the population is once again viewed as a resource which should not be wasted on emigration.

The element that triggers the growth process can be deduced from these various interpretations of development. A short-period increase in income can be straightforwardly achieved through growth in demand for locally produced goods and services. The latter takes the form of effective sectoral demand, also external to the local economy and possibly dynamic, which sets off a virtuous 'demand–supply' mechanism through Keynesian multiplier effects on income. In this case, the engine of development is demand. From this point of view, therefore, no consideration is made of the ability of supply to keep up with growing demand, given the assumption that there are no limits on local production capacity. But although this assumption may well be realistic in the short term, it is unsustainable in the long term. By contrast, if the focus is on individual well-being and long-term competitiveness, the engine of development must necessarily lie on the supply side, and specifically in the availability of production factors (labour, capital, entrepreneurship), and in the absolute and comparative advantages of the local firms which determine an area's production capacity and its position in the world market.

2.7 Theories of convergence and divergence: a distinction by now superseded

Handbooks on 'regional economics' have often drawn a distinction, indeed a dichotomy, between theories of convergence and divergence: that is, between theories which examine the reasons for diminishing disparities between rich and backward regions, and theories which, on the contrary, explain the persistence of those disparities.[28]

Ranged on the convergence side are theories originating within the neoclassical paradigm and which interpret (in their initial formulation) development as a process tending to equilibrium because of market forces. In equilibrium, not only is there an optimum allocation of resources but also an equal distribution of the production factors in space which guarantees, at least tendentially, the same level of development among regions.

On the divergence side stand theories of Keynesian origin which, by introducing positive and negative feedback mechanisms and the cumulative attraction and repulsion of productive resources respectively in a country's rich and poor areas, envisage not only the persistence but also the worsening of disparities among regions.[29]

In recent years, more refined mathematical and modelling tools have demonstrated that the same theories are able to explain both divergence and convergence. By introducing, for example, scale economies and agglomeration economies into a production function – obviously more complex than that of the 1960s model – the neoclassical model successfully simulates a series of behaviours and tendencies, both continuous and 'catastrophic', very distant from the mechanicism and univocity of the convergence predictions of the original neoclassical model. In the same way, the divergence yielded by Keynesian models (*à* la Myrdal and Kaldor in particular)[30] is called into question if the model's dynamic properties are analysed: according to the parameter values of the dynamic equations describing the model's economic logic, the local system either converges on a constant growth rate or explosively or implosively diverges from it.

It is therefore possible to conclude that there are no longer grounds for any dichotomy to be drawn between theories of convergence and divergence, between optimistic theories and pessimistic ones. However, the problem in and of itself is still very much present, and it is much more complex than was believed in the past. The neoclassical model, elegant in its formulation and consistent in its economic logic, has been frequently criticized as unsuited (in its original formulation) to interpretation of constant and persistent regional disparities. The Keynesian model, in its turn, has been faulted for being unable to foresee territorial limits to the evolution of the cumulative process, although these limits have substantial effects on territorial development paths. But if the 'theories of divergence–convergence' dichotomy is abandoned, the explanatory capacity of each theory can be recovered, to produce a broad array of conceptual tools with which to interpret the complex processes of territorial development.

2.8 Conclusions

The theories described in this chapter have highlighted the increasingly complex and intriguing ways in which models of economic growth treat space. The simple (and in certain respects banal) interpretation of space as uniform-abstract and straightforwardly relatable to administrative units – a space conceived as internally homogeneous and uniform, and which can therefore be synthesized into a vector of aggregate socio-economic-demographic features – has in recent years been replaced by a notion of diversified-relational space which restores to theories of regional development some of the founding principles of location theory: agglomeration economies and spatial interaction.

It is this more complex interpretation of space that has enabled regional economics to take decisive steps forward in analysis of local dynamics by conceiving space as the source of increasing returns and positive externalities. The development process also depends on the efficiency of the territorial organization of production, rather than solely on the quantity of economic resources available. Not only are the tangible elements of development (for example, the quantity of existing productive resources) important, so too are the intangible elements mentioned above: the learning processes, local relational networks and governance mechanisms that have increasing weight in defining an area's development path.

Finally, most recent years have seen an endeavour to escape from the impasse which caught regional economics between, on the one hand, growth theories of pure macroeconomic origin formalized into elegant models, and on the other, theories which abandon the rigour of formal treatment to consider new qualitative and territorial elements synthesizable – with due caution – into the concept of agglomeration economies. The most recent theories on local growth are able to incorporate increasing returns into the economic and formal logic of macroeconomics, and they are viewed (sometimes all too enthusiastically) as a new way to conceive space – as a means to merge previous conceptions together. Space is conceived as diversified; while territorial development is conceived as selective, cumulative and at increasing returns, and it is interpreted on the basis of a macroeconomic growth model.

It has been emphasized that this merger is in fact only an initially positive result. More detailed analysis shows that space is indeed conceived as diversified, but it receives no territorial explanation apart from one taking the form of the agglomera-

tion–non-agglomeration dichotomy. The territorial features (and the above-mentioned intangible elements) that play an important role in diversified-relational space theories by explaining and interpreting the level of competitiveness achieved disappear entirely in the macroeconomic models.

Still needed, therefore, is a convincing 'model' which comprises the micro-territorial, micro-behavioural and intangible elements of the development process. Required for this purpose is definition of patterns, indicators and analytical solutions to be incorporated into formalized models necessarily more abstract and synthetic in terms of their explanatory variables. A move in this direction is the quantitative sociology that embraces the paradigm of methodological individualism and seeks to 'measure' the social capital of local communities. It is obviously necessary to bring out territorial specificities within a macroeconomic model. Or, in other words, it is necessary to demonstrate the territorial microfoundations of macroeconomic growth models. This is the challenge facing regional economists in the years to come.

Notes

1. See Isard (1954, 1956).
2. See the well-known model of Dixit and Stiglitz (1977).
3. This chapter is mostly drawn on the introductory chapter of my textbook in regional economics. In that chapter for the first time I propose the distinction of space and its treatment in regional economic theories. See Capello (2007).
4. Among others, see von Thünen (1826), Hotelling (1929), Weber (1929 [1909]), Alonso (1960, 1964a), Christaller (1933 [1966]), Lösch (1954 [1940]).
5. Geographical (physical) features are removed from models and theories by assuming the existence of a homogeneous plain with equal fertility of land (Von Thünen, 1826) or uniform infrastructural endowment (Alonso, 1964b; Palander, 1935; Hoover, 1948; Christaller, 1933 [1966]; Lösch, 1954 [1940]).
6. See Hotelling (1929), Lösch (1954 [1940]).
7. See Von Thünen (1826), Alonso (1960, 1964a).
8. See Christaller (1933 [1966]), Lösch (1954 [1940]).
9. Ohlin defines a 'region' as a territory characterized by perfect mobility of production factors. See Ohlin (1933).
10. We refer here to the Keynesian regional growth theories of the 1950s (Hoyt, 1954; North, 1955); to the neoclassical interregional growth models (Borts, 1960 [1970]; Borts and Stein, 1964, 1968 [1962]); to the neoclassical interregional trade theory (Heckscher, 1919 [1950]; Ohlin, 1933).
11. The reference is, for example, to formalization of equilibrium in non-linearity conditions and equilibrium under monopolistic competition. The latter was proposed towards the end of the 1970s by Dixit and Stiglitz, and it provides the basis for some of the models.
12. See von Böventer (1975), p. 3. When von Böventer refers to ' "pure and exact" regional theory without agglomeration economies', he means the theories presented in Part II of this book; when he refers to ' "applied regional theory" which is inexact but takes agglomeration factors into account', he means theories expounded in more qualitative form, which will be the ones developed in this part of the book.
13. Many theories embrace the idea of an exogenous factor at the basis of regional development. For a dominant firm, see Perroux (1955); for the presence of infrastructure, see among others Aschauer (1989), Biehl (1991); for the spatial diffusion of innovation, see Hägerstrand (1967).
14. This is the case of the weak region achieving greater growth than the rich region in Borts and Stein's one-sector model. It must be stressed that the view of development adopted by other neoclassical models, like the Heckscher–Ohlin model, is one of generative development, not of competitive development. On the distinction between competitive and generative development see Richardson (1973, 1978).
15. See Perroux (1955), p. 308. For a critical re-examination of Perroux's theory, see Parr (1999a, 1999b).
16. See Camagni (2002).
17. Hence the name 'NEC areas'.
18. The 'industrial triangle' comprising Lombardy, Liguria and Piedmont, that is, the regions of north-western Italy.
19. See Ciciotti and Wettmann (1981), Johannisson and Spilling (1983), Stöhr and Tödtling (1977), Stöhr (1990), Secchi (1974), Garofoli (1981), Courlet and Pecqueur (1992), Ganne (1992). See Vásquez-Barquero (2002) for a well-structured survey of theories of endogenous development.

20. Becattini set out his main ideas in a study published in 1975 (see Becattini, 1975) and then developed them in a subsequent study of 1979 (see Becattini, 1979 [1989]). There followed a series of works in which Becattini expanded and deepened the concept of the 'Marshallian industrial district'. Recent volumes containing seminal works on the issue are Becattini (2004) and Brusco (1990).
21. See Marshall (1920). For detailed analysis of the links between Marshall's work and the theory of industrial districts see Bellandi (1989 [1982]).
22. Neo-Schumpeterian theories of local development are, among others, the milieu innovateur theory (Aydalot, 1986; Aydalot and Keeble, 1988; Camagni, 1991, 1999; Ratti et al., 1997; RERU, 1999); the learning region theory (Lundvall, 1992; Ludvall and Johnson, 1994; Maskell and Malmberg, 1999; Cooke, 2002). For a systematic review of neo-Schumpeterian local development theories, see Mouleart and Sekia (2003).
23. See Barentsen and Nijkamp (1989), Nijkamp and Reggiani (1988, 1992, 1993), Reggiani (2000). For an application to regional growth models, see Miyao (1981, 1984, 1987a, 1987b).
24. On the endogenous growth theories see among others, Romer (1986) and Lucas (1988); for a review, see Aghion and Howitt (1997). On the new economic geography, see Krugman (1991a, 1991b, 1991c), Krugman and Venables (1996), Fujita and Thisse (1996, 2002), Fujita et al. (1999); for a regional perspective, see Nijkamp and Poot (1998) and Nijkamp et al. (1998); for a critical survey, see Martin (1999).
25. In macroeconomics, the income multiplier effect is generated by the following process: an increase in one of the components of aggregate demand – for example demand for goods produced in the area (local consumption) – gives rise to a general increase in income. However, an increase in income in its turn generates an increase in consumption, and therefore in aggregate demand. The latter once again produces an increase in income, which once again generates increased consumption. The 'Keynesian multiplier' yields a value, by definition greater than unity, which measures the variation in output resulting from a unit change in some component of aggregate demand (consumption, investments, public spending, exports).
26. Note that per capita income as an indicator of disparity has the major shortcoming from the statistical point of view of associating better conditions of relative well-being with emigration from an area. In fact, increased per capita income is obtained either through real growth in regional income (the numerator in the income–population ratio) or through real growth in regional income (the denominator in the ratio). While the two effects are statistically recorded in the same way by the indicator, from the economic point of view they represent two very different cases: the former that of real economic growth; the latter that of possible social hardship and crisis.
27. The marginal productivity of a production factor, labour for example, measures the extent to which output varies with a change in one unit of labour. If the neoclassical law of decreasing marginal productivity holds, marginal productivity diminishes as the workforce of a firm (or an area) increases. Inevitably, therefore, surplus labour has nil marginal productivity. If new workers were included in the production process, they would be unable to produce additional units of output; for this reason, they remain unemployed.
28. The notion of 'backwardness' employed by regional economics should not be confused with the underdevelopment analysed by development economics. Although there are points of contact between the two disciplines – indeed, some of the early models of regional economics were decisively influenced by those of economic development theory – there are also important differences. The underdevelopment treated by regional economics is contextualized within a broader economic system (the country as a whole) with an already advanced level of industrialization on which backwardness can count: the 'Objective One' regions of the European Union, termed such because they have levels of per capita income below the average of European regions, are parts of economically advanced countries with infrastructure, technologies, labour forces and industrial systems typical of the industrialized world. The concern of development economics is instead with the underdevelopment of entire countries, and therefore also with the 'preconditions' for development: industrialization, population support, the creation of basic infrastructure and services for people and firms. Moreover, because regional economics deals with subnational territorial areas, it must disregard certain macroeconomic policy instruments, like the exchange rate or the interest rate, which belong among the public policy instruments available for country-level development.
29. See Meyer (1963), Isard (1956).
30. See Myrdal (1957) and Kaldor (1970).

References

Aghion, P. and P. Howitt (1997), *Endogenous Growth Theory*, Cambridge, MA: MIT Press.
Alonso, W. (1960), 'A theory of the urban land market', *Papers and Proceedings of the Regional Science Association*, **6**, 149–57.
Alonso, W. (1964a), *Location and Land Use: Towards a General Theory of Land Rent*, Cambridge, MA: Harvard University Press.

Alonso, W. (1964b), 'Location theory', in J. Friedmann and W. Alonso (eds), *Regional Development and Planning: A Reader*, Cambridge, MA: MIT Press, pp. 78–106.
Aschauer, D.A. (1989), 'Is public expenditure productive?', *Journal of Monetary Economics*, **23**, 177–200.
Aydalot, Ph. (ed.) (1986), *Milieux Innovateurs en Europe*, Paris: GREMI.
Aydalot, Ph. and D. Keeble (eds) (1988), *High Technology Industry and Innovative Environment*, London: Routledge.
Barentsen, W. and P. Nijkamp (1989), 'Modelling non-linear processes in time and space', in Å. Andersson, D. Batten, B. Johansson and P. Nijkamp (eds), *Advances in Spatial Theory and Dynamics*, Amsterdam: North-Holland, pp. 175–92.
Becattini, G. (1979), 'Dal Settore Industriale al Distretto Industriale. Alcune Considerazioni sull'Unità di Indagine dell'Economia Industriale', *Rivista di Economia e Politica Industriale*, **1**, 35–48; English edn. (1989), 'Sectors and/or districts: some remarks on the conceptual foundations of industrial economics', in E. Goodman and J. Bamford (eds), *Small Firms and Industrial Districts in Italy*, London: Routledge, pp. 123–35.
Becattini, G. (ed.) (2004), *Industrial Districts: A New Approach to Industrial Change*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.
Bellandi, M. (1989), 'The industrial district in Marshall', in E. Goodman and J. Bamford (eds), *Small Firms and Industrial Districts in Italy*, London: Routledge, pp. 136–52; orig. edn. (1982), 'Il Distretto Industriale in Alfred Marshall', *L'Industria*, **3**, July–September, 355–75.
Biehl, D. (1991), 'The role of infrastructure in regional development', in R.W. Vickerman (ed.), *Infrastructure and Regional Development*, London: Pion Limited, pp. 9–35.
Borts, G.H. (1960), 'The equalisation of returns and regional economic growth', *American Economic Review*, **50**, 319–47; reprinted in D. McKee, R. Dean and W. Leahy (eds) (1970), *Regional Economics: Theory and Practice*, New York: Free Press, pp. 147–76.
Borts, G.H. and J.L. Stein (1964), *Economic Growth in a Free Market*, New York: Columbia University Press.
Borts, G.H. and J.L. Stein (1968), 'Regional growth and maturity in the United States: a study of regional structural change', in L. Needleman (ed.), *Regional Analysis*, Harmondsworth: Penguin Books, pp. 159–97; orig. edn. (1962), in *Schweizerische Zeitschrift für Volkswirtscahft und Statistik*, **98**, 290–321.
Brusco, S. (1990), 'The idea of the industrial district: its genesis', in F. Pyke, G. Becattini and W. Sengenberger (eds), *Industrial Districts and Interfirm Cooperation in Italy*, Geneva: International Institute of Labour Studies, pp. 10–19.
Camagni, R. (1991), 'Local *milieu*, uncertainty and innovation networks: towards a new dynamic theory of economic space', in R. Camagni (ed.), *Innovation Networks: Spatial Perspectives*, London: Belhaven-Pinter, pp. 121–44.
Camagni, R. (1999), 'The city as a milieu: applying the GREMI approach to urban evolution', *Révue d'Economie Régionale et Urbaine*, **3**, 591–606.
Camagni, R. (2002), 'On the concept of territorial competitiveness: sound or misleading?' *Urban Studies*, **39** (13), 2395–2411.
Capello, R. (2007), *Regional Economics*, London: Routledge.
Christaller, W. (1933), *Die Zentralen Orte in Süddeutschland*, Darmstadt, Germany: Wissenschaftlische Buchgesellschaft; English edition (1966), *The Central Places in Southern Germany*, Englewood Cliffs, NJ: Prentice-Hall.
Ciciotti, E. and R. Wettmann (1981), 'The mobilisation of indigenous potential', Commission of the European Community, Internal Documentation on Regional Policy, n. 10.
Cooke, Ph. (2002), *Knowledge Economies: Clusters, Learning and Cooperative Advantage*, London: Routledge.
Courlet, C. and B. Pecqueur (1992), 'Les systémes industriels localisés en France: un nouvel model de developpement', in G. Benko and A. Lipietz (eds), *Les Régions qui Gagnent. Districts et Réseaux: les Nouveaux Paradigmes de la Géograqhie Economique*, Paris: Presses Universitaires de France, pp. 81–102.
Dixit, A. and J. Stiglitz (1977), 'Monopolistic competition and optimum product diversity', *American Economic Review*, **67** (3), 297–308.
Fujita, M. and J.-F. Thisse (1996), 'Economics of agglomeration', *Journal of the Japanese and International Economies*, **10**, 339–78.
Fujita, M. and J.-F. Thisse (2002), *Economics of Agglomeration: Cities, Industrial Location and Regional Growth*, Cambridge: Cambridge University Press.
Fujita, M., P. Krugman and A.J. Venables (1999), *The Spatial Economy: Cities, Regions and International Trade*, Cambridge, MA: MIT Press.
Ganne, B. (1992), 'Place et evolution des systèmes industriels locaux en France: economie politique d'une transformation', in G. Benko and A. Lipietz (eds), *Les Régions qui Gagnent. Districts et Réseaux: les Nouveaux Paradigmes de la Géograqhie Economique*, Paris: Presses Universitaires de France, pp. 315–45.
Garofoli, G. (1981), 'Lo Sviluppo delle Aree Periferiche nell'Economia Italiana degli Anni Settanta', *L'Industria*, **5** (3), 391–404.

Hägerstrand, T. (1967), *Innovation Diffusion as a Spatial Process*, Chicago, IL: University of Chicago Press.
Heckscher, E.F. (1919), 'The effect of foreign trade on the distribution of income', *Economisk Tidskrift,* **21**, 497–512; reprinted in H.S. Ellis and L.A. Metzler (eds) (1950), *Readings in the Theory of International Trade*, Allen & Unwin, London, pp. 270–300.
Hoover, E.M. (1948), *The Location of Economic Activity*, New York: McGraw-Hill.
Hotelling, H. (1929), 'Stability in competition', *Economic Journal*, **39** (153), 41–57.
Hoyt, H. (1954), 'Homer Hoyt on the development of economic base concept', *Land Economics*, May, 182–7.
Isard, W. (1954), 'Location theory and trade theory: short run analysis', *Quarterly Journal of Economics*, **68** (2), 305–20.
Isard, W. (1956), *Location and Space-Economy*, Cambridge, MA: MIT Press.
Johannisson, B. and O. Spilling (1983), *Strategies for Local and Regional Self-Development*, Oslo: NordREFO.
Kaldor, N. (1970), 'The case of regional policies', *Scottish Journal of Political Economy*, **3**, 337–48.
Krugman, P. (1991a), *Geography and Trade*, Cambridge, MA: MIT Press.
Krugman, P. (1991b), 'Increasing returns and economic geography', *Journal of Political Economy*, **99** (3), 484–99.
Krugman, P. (1991c), 'History vs. expectations', *Quarterly Journal of Economics*, May, 651–67.
Krugman, P. and A.J. Venables (1996), 'Integration, specialisation and adjustment', *European Economic Review*, **40**, 959–67.
Lösch, A. (1954), *The Economics of Location*, New Haven, CT: Yale University Press; orig. edn (1940), *Die Räumlische Ordnung der Wirtschaft*, Jena: Gustav Fischer.
Lucas, R. (1988), 'On the mechanics of economic development', *Journal of Monetary Economics*, **22**, 3–42.
Lundvall, B.-A. (1992), 'Introduction', in B.-A. Lundvall (ed.), *National Systems of Innovation. Towards a Theory of Innovation and Interactive Learning*, London: Pinter Publishers, pp. 1–19.
Lundvall, B.-A. and B. Johnson (1994), 'The learning economy', *Journal of Industry Studies*, **1** (2), 23–42.
Marshall, A. (1920), *Principles of Economics*, 8th edn, London: Macmillan.
Martin, R. (1999), 'The new "geographical turn" in economics: some critical reflections', *Cambridge Journal of Economics*, **23**, 65–91.
Maskell, P. and A. Malmberg (1999), 'Localised learning and industrial competitiveness', *Cambridge Journal of Economics* **23** (2), 167–85.
Meyer, J.R. (1963), 'Regional economics: a survey', *American Economic Review*, **53**, 19–54.
Miyao, T. (1981), *Dynamic Analysis of the Urban Economy*, New York: Academic Press.
Miyao, T. (1984), 'Dynamic models of urban growth and decay: a survey and extensions', paper presented at the Second World Conferences of Arts and Sciences, Rotterdam, 4–15 June.
Miyao, T. (1987a), 'Dynamic urban models', in E. Mills (ed.), *Urban Economics: Handbook of Regional and Urban Economics*, Amsterdam: North-Holland, Vol. 2, pp. 877–925.
Miyao, T. (1987b), 'Urban growth and dynamics', in T. Miyao and Y. Kanemoto (eds), *Urban Dynamics and Urban Externalities*, Chur, Switzerland and New York: Harwood Academic Publishers, pp. 1–41.
Mouleart, F. and F. Sekia (2003), 'Territorial innovation models: a critical survey', *Regional Studies*, **37** (3), 289–302.
Myrdal, G. (1957), *Economic Theory of Under-developed Regions*, London: General Duckworth & Co.
Nijkamp, P. and J. Poot (1998), 'Spatial perspectives on new theories of economic growth', *Annals of Regional Science*, **32** (1), 7–38.
Nijkamp, P. and A. Reggiani (1988), 'Entropy, spatial interaction models and discrete choice analysis: static and dynamic analogies', *European Journal of Operational Research*, **36**, 186–96.
Nijkamp, P. and A. Reggiani (1992), *Interaction, Evolution and Chaos in Space*, Berlin: Springer Verlag.
Nijkamp, P. and A. Reggiani (1993), *Non-linear Evolution of Spatial Economic Systems*, Berlin: Springer-Verlag.
Nijkamp, P., R. Stough and E. Verhoef (eds) (1998), 'Endogenous growth in a regional context', *Annals of Regional Science*, **32** (1), special issue.
North, D. (1955), 'Location theory and regional economic growth', *Journal of Political Economy*, **63**, 243–58.
Ohlin, B. (1933), *Interregional and International Trade*, Cambridge, MA: Harvard University Press.
Palander, T. (1935), *Beitrage zur Standortstheorie*, Uppsala: Almqvist & Wiksells Boktryckeri.
Parr, J. (1999a), 'Growth pole strategies in regional economic planning: a retrospective view. Part 1: origins and advocacy', *Urban Studies*, **36** (7), 1195–1216.
Parr, J. (1999b), 'Growth pole strategies in regional economic planning: a retrospective view. Part 2: implementation and outcome', *Urban Studies*, **36** (8), 1247–68.
Perroux, F. (1955), 'Note sur la notion de pôle de croissance', *Economie Appliquée*, **7** (1–2), 307–20.
Ratti, R., A. Bramanti and R. Gordon (eds) (1997), *The Dynamics of Innovative Regions*, Aldershot: Ashgate.
Reggiani, A. (2000), 'Introduction: new frontiers in modelling spatial and economic systems', in A. Reggiani (ed.), *Spatial Economic Science*, Berlin: Springer-Verlag, pp. 1–11.
RERU (1999), 'Le paradigme du milieu innovateur dans l'economie contemporaine', *Revue d'Economie Régionale et Urbaine*, **3**, special issue.

Richardson, H.W. (1973), *Regional Growth Theory*, London: Macmillan.
Richardson, H.W. (1978), *Regional and Urban Economics*, Harmondsworth: Penguin Books.
Romer, P. (1986), 'Increasing returns and long-run growth', *Journal of Political Economy*, **94** (5), 1002–37.
Secchi, B. (1974), *Squilibri Regionali e Sviluppo Economico*, Padova, Italy: Marsilio.
Stöhr, W. (1990), 'On the theory and practice of local development in Europe', in W. Stöhr (ed.), *Global Challenge and Local Responses*, London: Mansell Publishing, pp. 35–54.
Stöhr, W. and F. Tödtling (1977), 'Spatial equity: some anti-thesis to current regional development doctrine', *Papers of the Regional Science Association*, **38**, 33–53.
Vásquez-Barquero, A. (2002), *Endogenous Development*, London: Routledge.
von Böventer, E. (1975), 'Regional growth theory', *Urban Studies*, **12**, 1–29.
von Thünen, J.H. (1826), *Der Isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*, Hamburg: Puthes.
Weber, A. (1929), *Alfred Weber's Theory of the Location of Industries*, Chicago, IL: University of Chicago Press; orig. edn (1909), *Über der Standort der Industrien*, Tübingen: Verlag Mohr.

3 Location/allocation of regional growth

Gunther Maier and Michaela Trippl

3.1 Introduction

In this chapter we look at the spatial distribution of regional growth. The main goal is to analyse the implications that various theories of regional growth have for the spatial distribution of economic activities and the long-term dynamics of the regional economy. Two views will be considered: first, the view of the neoclassical model; second, that of endogenous growth theory and new economic geography. The neoclassical model is briefly presented in section 3.2 of the chapter. Section 3.3 sketches the main features of endogenous growth theory and new economic geography. We will argue that endogenous growth theory and new economic geography apply the same basic logic, namely to introduce externalities into a general equilibrium model. In our view, the introduction of externalities, which according to endogenous growth theory is necessary in order to understand long-term growth processes, is the main innovation of the new theories. The implications of this step, however, are dramatic. We discuss them in sections 3.4 and 3.5 of the chapter. Briefly speaking, externalities lead to non-linearities in the growth process which may generate complex system dynamics including chaotic behaviour. These models call into question most of the results of the neoclassical theory with the corresponding consequences for economic policy. Conclusions are drawn in section 3.6 of the chapter.

3.2 The allocation of growth in the neoclassical model

The standard neoclassical model of regional growth serves as a reference model in our discussion. It borrows key elements from the neoclassical growth theory (Solow, 1956; Swan, 1956) and applies them to the regional context. The neoclassical model of regional growth is based upon the standard assumptions of neoclassical economics: utility maximization, perfect mobility, perfect information and perfect competition. One specific aspect of the assumption of perfect competition is of particular relevance in our context. This is the assumption of a linear homogeneous production function. A production function

$$Y = F(K, L) \tag{3.1}$$

is linear homogeneous when $F(0,0) = 0$ and $F(aK,aL) = aF(K,L)$ for $a > 0$ and all values of capital (K) and labour (L). Such a production function exhibits constant returns to scale. Increasing or reducing the level of production does not change the efficiency of the production process. This precludes any fixed costs in the production process. A specific version of a linear homogeneous production function is the Cobb–Douglas production function $Y = AK^{\alpha}L^{1-\alpha}$. To incorporate technological progress, the function can be specified as

$$Y_t = Ae^{\lambda t}K^{\alpha}L^{1-\alpha} \tag{3.2}$$

Technical knowledge is assumed to increase with a factor λ per time period. In this way it shifts up the function and determines the long-run growth rate of the economy. We will further discuss this relationship in the following section.

Perfect competition also implies an atomistic market structure and precludes any externalities. No actor is so important in the market that he or she can influence the price strategically. The behaviour of the actors has no side-effects on other actors or other parts of the system besides those via the aggregate market. Externalities will play a crucial role in the latter part of our discussion.

The neoclassical assumptions have a number of important implications. First, they imply equilibrium in all markets. If any market were in disequilibrium, prices would change and the perfectly informed and utility-maximizing actors would react accordingly. Second, production factors are paid the value of their marginal product. If wages, for example, were lower, labour input would be lower than optimal and firms could increase production and their profits by paying higher wages. If they were too high, labour costs would be too high and reduce profits. Consider two regions with identical production functions and identical levels of labour. The two regions differ by their initial level of capital. Region *p*, the poor region, has a low initial level of capital, region *r*, the rich region, a high initial level of capital ($K_p < K_r$). From equation (3.1) it is immediately clear that this implies $Y_p < Y_r$, verifying the characterization of the regions.

The level of capital is determined endogenously in the model. In every period a given percentage, *s*, of production is saved and reinvested into the economy. On the other hand, a fixed percentage, δ, of the existing capital stock becomes obsolete and is depreciated. So, from one period to the next in every region the stock of capital grows according to

$$K_{it} = K_{it-1} + s\,Y_{it-1} - \delta K_{it-1}$$

When we solve this difference equation, we find that in both regions capital and consequently also output converges toward the same steady state level. Also, we see that the growth rate of output is higher in the poor region than in the rich one (see Figure 3.1). Note that we have made no statements about any relation between the two regions. In fact, they can be completely isolated from each other. Because of the process of capital accumulation, they grow such that the poor region catches up and both reach the same output level in the long run. The main reason for this is the shape of the production function.

Although capital accumulation is sufficient for convergence in the neoclassical model, the neoclassical model of regional growth is typically associated with interregional flows of resources. This view goes back to Borts and Stein (1964). Since production factors are paid according to the value of their marginal product, as we have argued above, wages and capital rents differ between the rich and poor regions. Since the poor region lacks capital as compared to the rich one, capital rents, *r*, are higher and wages, *w*, lower there:

$$r_r = \frac{\partial Y_r}{\partial K_r} < r_p = \frac{\partial Y_p}{\partial K_p};\quad w_r = \frac{\partial Y_r}{\partial L_r} > w_p = \frac{\partial Y_p}{\partial L_p}$$

Since all actors are utility-maximizers and perfectly informed, capital will flow from the rich to the poor and labour from the poor to the rich region. This will equilibrate the capital–labour ratio ($k = K/L$) between the regions as well as capital rent and wages. This

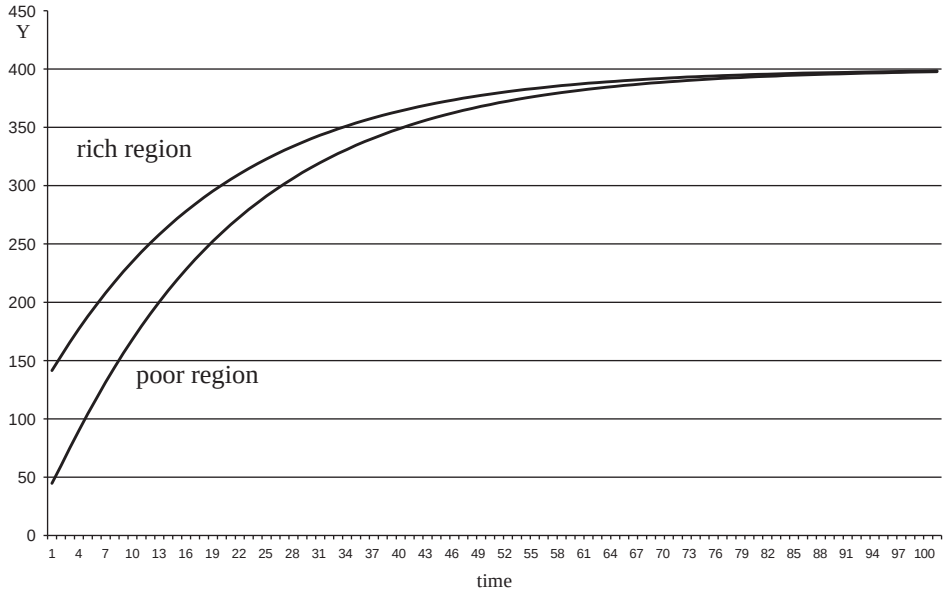


Figure 3.1 Growth and convergence due to capital accumulation in the neoclassical model

mechanism is sketched in Figure 3.2. Since capital and labour are used more efficiently after the convergence process, the total output in both regions together increases.

The neoclassical model of regional growth has clear predictions and policy implications. The growth and convergence process leads to an even distribution of per capita income and equilibrates wages and capital rents. Irrespective of the initial amount of capital in a region, growth always leads to the same steady state in the long run. Any disturbances of the process are eliminated over time. In this sense, history does not matter in this model and the long-term outcome is perfectly predictable. It can also be shown that in an economy where all the neoclassical assumptions hold, the long-term outcome is Pareto optimal. These results imply that policy has no major role to play as the long-term outcome is optimal and reached automatically. The underlying message that the economy should develop freely and not be disturbed by policy has been repeated frequently in theoretical literature as well as in policy strategies and documents. As we shall see below, contemporary views of regional growth offer much less support for this position.

3.3 Endogenous growth and new economic geography

While the neoclassical model provides a consistent view of regional growth with clear implications and policy recommendations, some of its implications are rather disturbing. The first one is related to the steady state solution, where all growth due to capital accumulation comes to an end. When we take into account innovation as in equation (3.2), the economy will grow in the long run with the same rate as technical knowledge, to which innovation adds. But the growth of technical knowledge cannot be explained within the neoclassical model and has to remain exogenous. This leads to the unsatisfactory result that in the long run the neoclassical growth model explains growth by something that remains unexplained. This has led to the development of endogenous growth models in the 1980s and 1990s.

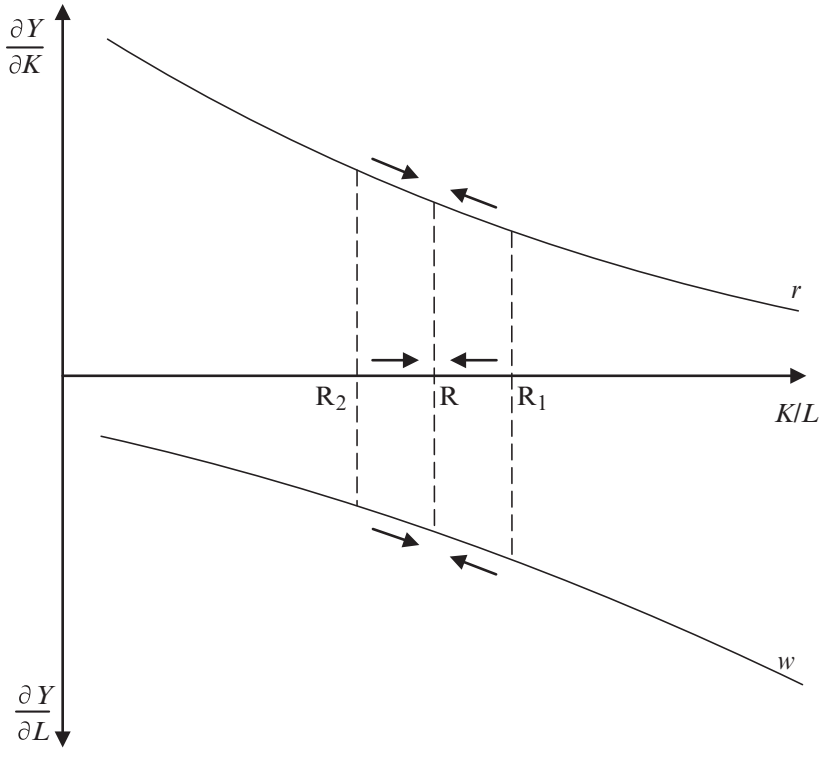


Figure 3.2 Convergence due to factor mobility in the neoclassical model

Another disturbing implication becomes obvious from a regional perspective. Consider a number of regional economies, each operating according to the neoclassical model. When we allow for transport costs for the shipment of goods between the regions, in equilibrium there will be no goods transported between the regions. Utility-maximizing producers would only produce in one region for consumption in another, when they would have a high enough benefit from concentrating production so that they could overcompensate for the implied transportation costs. However, a linear homogeneous production function does not offer any advantages of concentrated production. Therefore, in the neoclassical model no producer would take this option. This result has been claimed by Mills (1972) and formally shown by Starrett (1978). When we break down the regions into smaller and smaller areas, in the limit we get a result that is adequately called 'backyard capitalism' (Fujita et al., 1999). All products that a household consumes are produced right in its backyard in order to avoid transportation costs. With the same line of reasoning we can argue that the only workers in this production will be the members of the respective household. 'Each consumer becomes a Robinson Crusoe producing for his own consumption' (Ottaviano and Puga, 1997, p. 3).

The endogenous growth and the new economic geography literature attempts to overcome these conceptual problems. In an early attempt, Romer (1986) modelled technological progress as 'the accidental, and indirect, outcome of decisions to invest in capital accumulation' (Angeriz et al., 2006, p. 3), introducing an externality into the growth

process. This version of the endogenous growth model is therefore called the 'externality model' (Bröcker, 1994). Latter contributions (for example Romer, 1990; Grossman and Helpman, 1991) explicitly introduce a sector that produces new technologies. In order to have an incentive for this production, this sector typically is assumed to have a monopoly over these new technologies and thus enjoys a monopoly rent. 'In both cases, fundamental to the story of endogenous growth is the existence of knowledge spillovers, leading to the existence of increasing returns, as, without increasing returns, growth would dry up in the absence of an exogenous driving force' (Angeriz et al., 2006, p. 3). Or, formulated differently, it turned out that one can only explain growth endogenously when one departs from the neoclassical assumptions and allows for a mechanism that generates increasing returns to scale.

In a spatial context, increasing returns to scale imply positive agglomeration effects: concentrating production in one economic area allows for higher productivity. The implications of these effects on the spatial distribution of activities are analysed by the new economic geography literature (Krugman, 1991a, 1991b; Ottaviano and Puga, 1997; Fujita et al., 1999). A monopolistically competitive sector (Dixit and Stiglitz, 1977) tends to agglomerate economic activity, while transport costs and an immobile sector tend to pull it apart. These counteracting effects of agglomeration benefits on the one hand and transportation costs on the other are at the heart of all new economic geography models. Venables (2006) argues that despite different structures, arguments and spatial scope, all new economic geography models 'require two building blocks. One is an understanding of the costs of distance, and the other is a description of the mechanisms that cause activity to cluster' (p. 740).

The new economic geography literature emphasizes the spatial consequences of the agglomeration forces that are necessary in order to understand endogenous growth. It shows that spatial disparities at different spatial levels may develop endogenously from the economic processes and that therefore 'spatial disparities are a normal economic outcome' (Venables, 2006, p. 751). One of the first new economic geography models, Krugman's core–periphery model, illustrates this outcome and the basic mechanisms leading to it quite clearly. Similar results have been shown at the levels of cities, systems of cities, and countries. The introduction of agglomeration forces has also rejuvenated some older literature and discussion that has argued for polarizing effects in economic processes as they tend to amplify those effects rather than dampen them as in the neoclassical world. With agglomeration forces at work, a small change in the location pattern, say one firm or one worker moving from one region to another, may offset a cumulative process that more and more concentrates economic activities in the one region. Examples of such older literature are Marshall (1920), Myrdal (1957) and Hirschman (1958). They have made arguments in favour of concentration tendencies in economics, but could not integrate them in a general equilibrium framework as the new economic geography literature does. However, in the light of the achievements of new economic geography their arguments and those of many others who focus on specific parts of the network of economic relations can gain new relevance.

The arguments that endogenous growth theory and new economic geography introduce into the traditional general equilibrium framework lead to a view of the economy which differs markedly from that of the neoclassical model. These arguments can be illustrated by use of the so-called tomahawk bifurcation which is sketched in Figure 3.3.

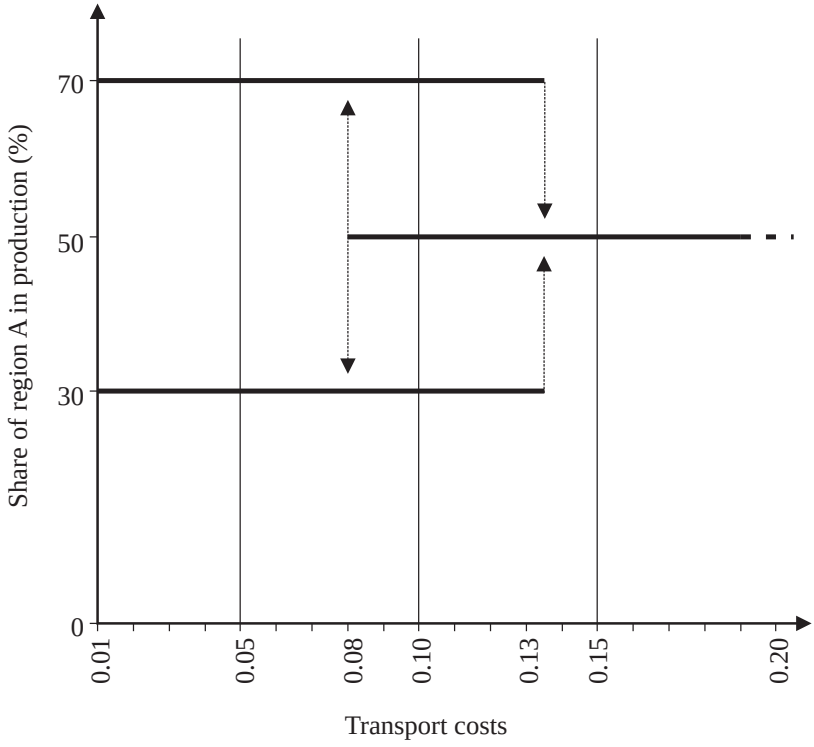


Figure 3.3 Tomahawk bifurcation

The figure shows the equilibrium share of one region (region A) of total production at various levels of transport costs in the two-region core–periphery model of Krugman (1991a). Our first observation is that the model can yield multiple equilibria. While at transport costs of 0.15 the only equilibrium is an equal distribution between the regions, at transport costs of 0.05, region A can either have 30 per cent or 70 per cent of production. At transport costs of 0.10, all three equilibria are possible: 30 per cent, 50 per cent and 70 per cent. All these equilibria are stable in the sense that no economic agent has an incentive to move to the other region, once the system has reached the equilibrium. The second observation is path-dependence. Suppose the system has reached a 50 per cent equilibrium under transport costs of 0.15. When transport costs fall to 0.10, the equal distribution will pertain, because it is a stable equilibrium. Only when transport costs fall below 0.08 will the 50 per cent equilibrium become unstable and region A will either drop to 30 per cent of production or rise to 70 per cent. Which of the two equilibria will be reached may depend upon small random events. However, when we start from transport costs of 0.05 and a production share of 30 per cent of region A, the region will not increase its share of production when transport costs increase to 0.10. They will remain at the stable 30 per cent equilibrium. Only at transport costs of 0.13 or higher will the system switch to an equal distribution. Our third observation is lock-in. In the situation of increasing transport costs and a 30 per cent equilibrium, a policy incentive that moves the distribution toward equal shares will fail when it is not

substantial enough. The system will be locked in at the current share and return to the stable 30 per cent equilibrium.

3.4 Agglomeration factors as externalities

All these peculiar features that we have discussed above result mainly from the fact that new economic geography introduces agglomeration forces into the general equilibrium framework. Transport costs just keep the system from collapsing into one location. Agglomeration forces introduce an extra relation between economic agents. The availability of other agents at a location or region allows producers to produce more efficiently. Since the availability of other agents follows from their location decisions, agglomeration forces also introduce an extra linkage between the location decisions of actors over time.

Because it applies a general equilibrium framework, new economic geography can handle agglomeration forces only in a rather stylized form. Other, more partial approaches have put forward additional arguments in favour of the existence of agglomeration forces in the economy.

One of the first who argued in favour of spatial proximity and agglomeration was Alfred Marshall (1920). He described three mechanisms. First, the linkages between firms in the value chain: spatial proximity between firms that are in a buyer–supplier relation allows them to save on transport costs and enjoy economies of scale. Second, Marshall argues that there is a thick labour market in agglomerated areas. For employers the agglomeration offers the advantage of a good choice of workers with the required skills. For workers a concentration of potential employers allows them to specialize and to develop and utilize sector-specific knowledge. The third argument by Marshall is that of technological externalities. 'Good work is rightly appreciated, inventions and improvements in machinery, in processes and the general organization of the business have their merits promptly discussed; if one man starts a new idea, it is taken up by others and combined with suggestions of their own; and thus it becomes the source of further new ideas' (Marshall, 1920, p. 225).

Marshall's third argument in particular has gained substantial support both theoretically and empirically in recent years. Knowledge spillovers are seen as having a positive impact on innovation and growth (Maier and Sedlacek, 2005; Döring and Schnellenbach, 2006; Lim, 2007). Conceptualized as a key explanatory factor for the spatial concentration of production and innovation activities, they are at the heart of modern cluster theories (Keeble and Wilkinson, 2000; Malmberg and Maskell, 2002, 2006) and the regional innovation system approach (Acs, 2000; Asheim and Gertler, 2005). The theory on regional innovation systems states that knowledge creation is a path-dependent process and considers an intense local knowledge circulation as a critical condition for a high-innovation performance of regions. Applying a knowledge production function approach Audretsch and Feldman (1996), Bottazzi and Peri (2003) and others demonstrate empirically the importance of local knowledge spillovers.

3.5 Implications of externalities

In a technical sense all the arguments of endogenous growth theory and of the new economic geography, and of the more partial views, introduce externalities into the economic system. Through unintended side-effects of their decisions, economic agents indirectly influence the decisions of others.

Efficiency, feedback loops and switching regimes

In the neoclassical model externalities are seen as isolated phenomena that disturb the market process. Since these effects are not taken into account by the economic agents when making decisions, externalities typically lead to inefficient outcomes. Consequently, in a neoclassical view policy should attempt to internalize the respective externality. The hypothetical outcome of the economy under neoclassical assumptions – without this externality – serves as a yardstick for calculating the tax or subsidy necessary for internalization (Mishan, 1971; Lin, 1976). This requires, however, that no other externality besides the one under investigation exists in the economy. When this condition does not hold, 'we know from the General Theory of the Second Best (Lipsey and Lancaster, 1957) that there is no certainty that the measure taken to internalize the one externality will move the economy as a whole closer to a Pareto optimum' (Maier and Sedlacek, 2005, p. 5). The result with the one externality internalized may be worse than before.

The contemporary view of endogenous growth theory and new economic geography departs radically from the neoclassical perspective on externalities. As Venables (2006) summarizes: 'this view of the world suggests that externalities . . . are all pervasive' (p. 751). It seems that spatial proximity, knowledge production and innovation, network linkages, infrastructure, environmental effects (Johansson and Quigley, 2004; Batabyal and Nijkamp, 2004; McCann and Shefer, 2004) and many other relationships add up to a tissue of externalities that spans the economy. This multitude of unintended side-effects yields a dynamic system where the structure that exists at a certain point in time influences the forces that advance the system over time. Such a non-linear feedback loop may produce highly complex dynamics of the system.

The importance of externalities for the long-run behaviour of a dynamic system is nicely illustrated by Arthur et al. (1987). They compare two versions of a very simple dynamic model. In every period one unit is added to one of two containers by a random process. We can think of the containers as regions and of the units as economic activities. In the first version of the model the probability that the activity is assigned to region A is fixed and constant over time. Because of the law of large numbers, in the long run the region's share of economic activity will approach the respective probability. There is one long-run equilibrium that we can predict with certainty. If the process is disturbed by some exogenous event, once it has ended, its impact will be eliminated over time by the growth process. The dynamic behaviour of this version of the model corresponds to that of the neoclassical model that we have discussed in section 3.2.

In the second version, an externality is introduced into the model. Instead of being constant and exogenously given, the probability that the next economic activity is assigned to region A is assumed to be equal to this region's current share of economic activity. The outcome of the assignment process at one time period changes not only the distribution of economic activities at that period, but also the chance of the future assignment processes. The result is a so-called Polya process. Polya (1931) showed that in the long run the relative frequencies resulting from this process tend toward a limit X with probability one, where 'X is a random variable uniformly distributed between 0 and 1' (Arthur, 1994, p. 36). In other words, the distribution of economic activities will converge toward equilibrium, but there are infinitely many equilibria possible and at the beginning of the process all are equally likely. At closer inspection we see that the second version of the model, that is, the version with the externality, shows similar characteristics to the

endogenous growth and new economic geography models: multiple equilibria, path-dependence and lock-in. Since there is an infinite number of equally likely equilibria, the long-term outcome of the process is completely unpredictable at its beginning. Although the two versions of the model differ only by the added externality, their dynamics and long-term results differ considerably.

This version of Arthur's model is rather abstract. In this form it demonstrates clearly the dramatic consequences externalities may have. The underlying ideas can easily be used in a more economic context. Arthur (1994) applied them to competition between new technologies, industrial location and the transmission of information. Maier (2001) combines the same logic with a two-region neoclassical model of the form described in section 3.2. He assumes that capital is perfectly mobile, labour immobile between the regions. For the innovation process he applies the logic of Arthur's model. Instead of being accumulated simultaneously in both regions over time, Maier (2001) assumes that in every time period one unit of innovation is added to one of the regions, where the assignment is random and exactly as in Arthur's model. When the assignment probability is constant, exogenously given, and identical for the two regions, the model behaves like the neoclassical growth model discussed in section 3.2. However, when the assignment probability is set equal to the region's share of production, the model's dynamic behaviour changes dramatically. Up to a certain time period the distribution of production between the two regions tends toward equal shares. Past this time period, however, the equal distribution equilibrium becomes unstable and production starts to accumulate in one of the regions. The other region's share of production will in the long run tend toward zero, despite the fact that this region has the same number of workers and that wages are flexible and decline accordingly.

This option of switching equilibria raises an obvious policy dilemma. Since the system behaves like the neoclassical model in the early phase, policy-makers may believe that it will converge to equal shares also in the long run, and interpret deviations from the equal distribution as temporary even after the equilibrium has switched. Therefore, policy may not intervene and the region will become locked into an undesirable development path.

Externalities, growth and chaos

The above-mentioned link between the current structure of the system and its temporal adjustment process that is introduced through externalities opens up another set of disturbing possibilities. Such non-linear feedback loops are at the heart of all models that produce complex dynamics including chaos (Puu, 1997, 2000). A well-known example is the logistic function $y_{t+1} = ay_t(1 - y_t)$. In economic terms, this function can be interpreted as a growth model with an externality in form of a capacity constraint. Let y be output (Y) relative to some upper limit (M), that is, $y = Y/M$. Then, $(1 - y_t)$ measures how close output is to the upper limit at time t. Since the rest of the equation, $y_{t+1} = ay_t$ represents a standard exponential growth process, we can consider the term in parentheses as a factor that dampens growth, when the system approaches the upper limit. The growth rate, y_{t+1}/y_t, turns out to be $a(1 - y_t)$ and to depend upon the level of output at time t.

As discussed, for example, in Peitgen et al. (1998), despite its simplicity the logistic function is able to generate the full range of non-linear dynamics: sensitive dependence on initial conditions, period doubling, chaotic regimes with embedded windows of stability,

and so on. Baumol and Benhabib (1989) use the logistic function in their early account of chaotic dynamics in economic models. Day (1982) embeds the logic of the logistic function into 'the familiar, neoclassical theory of capital accumulation' (p. 406) to produce examples of chaotic trajectories. Currie and Kubin (1995) derive the logistic function from a simple model of two markets with a production lag.

The amount of literature discussing complex non-linear dynamics in economics has grown rapidly in recent years. We will just mention a few recent examples. Currie and Kubin (2006) apply Krugman's core–periphery model to produce chaos, Gomes (2007) uses a Solow-like growth model with migration and a congestion externality for the same result. Auray et al. (2002) base their story of chaotic behaviour on a monetary model while Yousefi et al. (2000) generate chaos from a model of interdependent open economies.

The common denominator of all this literature is the externality that introduces a non-linearity into the growth process. Because of that, small disturbances resulting, for example, from a measurement error, some rounding of parameter values or some external influence, may be amplified over time and eventually dominate the trajectory of the system. While the general acceptance of externalities as an essential part of the economy calls for a more active role of policy than suggested by the neoclassical view (Venables, 2006), the new models can hardly produce any policy guidelines. Obviously, a chaotic system is unsuitable as a policy target. But even when the system is non-chaotic, because of the externalities, its long-term outcome may be inefficient. Depending on the state of the system, the same policy intervention may either dramatically change the trajectory of the process or have no long-term impact at all. The same holds for exogenous or indirect disturbances. An exogenous change or an indirect side-effect of the process via the environment or the society, for example, however small, may lead to a completely different long-term outcome. The fundamental postulate of economic policy, the 'principle of the negligibility of indirect effects' (Schumpeter, 1954, p. 990), does not necessarily hold in the non-linear world of an economy with externalities.

3.6 Summary and conclusions

In this chapter we have reviewed recent developments in the theoretical literature on regional growth and argue that all this literature suggests that externalities are an essential element of the economy which cannot be ignored. The endogenous growth theory even shows that without externalities the long-term growth process cannot be explained.

This is a fundamental deviation from the traditional neoclassical view of economics that is still dominating our thinking. The neoclassical model, which we sketched in section 3.2, has a well-defined long-term result in terms of the distribution of output and growth, wages and capital rent. It also generates clear policy recommendations based on its result about the efficiency of market processes. These results have been used extensively by researchers, consultants and policy-makers alike in suggesting or designing economic policies.

However, some of the disturbing implications of the neoclassical model have led to the above-mentioned change in perspective. From a regional point of view, the new economic geography appears to be the most important theoretical element of this paradigm shift. We sketched endogenous growth theory and new economic geography in section 3.3 of this chapter. As we argue, the essential step in this theory is the introduction of agglomeration forces – counteracting transport costs – which may lead to regional differences in

the distribution of economic activities. The basis of the agglomeration forces are externalities between the actors and elements of the economic system.

By accepting externalities as an important element of the economy as suggested by the new theories – otherwise neither long-term growth nor spatial structures can be understood – we open a kind of Pandora's box and release various phenomena that are unknown to the neoclassical model (section 5): multiple equilibria, path-dependence and lock-in, sensitivity to initial conditions, small disturbances and indirect effects, sensitivity to marginal changes in parameters, chaotic behaviour and convergence toward strange attractors. Obviously, by accepting externalities, the new theories take a radically new view of the economy. However, what the consequences and implications of this new view are is by no means clear yet. As far as policy is concerned, the new theories can provide much less guidance than the neoclassical model. Statements about automatic tendencies toward equilibrium or convergence, about efficient results of the market process, about the negligibility of small disturbances, and so on, are generally not justifiable under the new theories. This does not mean that we should stay away from policy. To the contrary, because of possible side-effects of economic processes, path-dependence and lock-in policy will have to try to correct negative developments (see Baldwin et al., 2003 for a discussion of public policy in this context). But designing policy appears to be much more difficult under the new theories. 'In the details of . . . policy prescriptions these models open up a Pandora's box of contradictions' (Brakman and van Marrewijk, 1996, p. 252). The externalities that we allow to enter our theories challenge the simple prescriptions of neoclassical economics. As far as theory is concerned, 'many new questions await to be answered and some old questions need to be reconsidered' (Maier and Sedlacek, 2005, p. 13). Although the step toward externalities seems unavoidable, it is not without risks. In 1939 Hicks 'drew back in alarm' (Arthur, 1994, p. 4) when he surveyed the possibilities of departure from the assumptions of perfect competition. 'The threatened wreckage is that of the greater part of economic theory' (Hicks, 1939, p. 84).

References

Acs, Z. (ed.) (2000), *Regional Innovation, Knowledge and Global Change*, London: Pinter.

Angeriz, A., J. McCombie and M. Roberts (2006), 'Some new estimates of returns to scale for EU regional manufacturing, 1986–2002', CCEPP, Working Papers, WP03-06.

Arthur, W.B. (1994), *Increasing Returns and Path Dependence in the Economy*, Ann Arbor, MI: University of Michigan Press.

Arthur, W.B., Y.M. Ermoliev and Y.M. Kaniovski (1987), 'Path-dependent processes and the emergence of macrostructure', *European Journal of Operational Research*, **30**, 294–303.

Asheim, B. and M. Gertler (2005), 'The geography of innovation', in J. Fagerberg, D. Mowery and R. Nelson (eds), *The Oxford Handbook of Innovation*. Oxford: Oxford University Press, pp. 291–317.

Audretsch, D. and M. Feldman (1996), 'Innovative clusters and the industry life cycle', *Review of Industrial Organisation*, **11**, 253–73.

Auray, S., F. Collard and P. Fève (2002), 'Money and external habit persistence: a tale of chaos', *Economic Letters*, **76**, 121–7.

Baldwin, R.E., R. Forslid, P. Martin, G. Ottaviano and F. Robert-Nicoud (2003), *Economic Geography and Public Policy*, Princeton, NJ: Princeton University Press.

Batabyal, A.A. and P. Nijkamp (2004), 'The environment in regional science: an eclectic review', *Papers in Regional Science*, **83**, 291–316.

Baumol, W.J. and J. Benhabib (1989), 'Chaos: significance, mechanism, and economic applications', *Journal of Economic Perspectives*, **3**, 77–105.

Borts, G.H. and J.L. Stein (1964), *Economic Growth in a Free Market*, New York: Columbia University Press.

Bottazzi, L. and G. Peri (2003), 'Innovation and spillovers in regions: evidence from European patent data', *European Economic Review*, **47**, 687–710.

Brakman, S. and Ch. van Marrewijk (1996), 'Trade policy under imperfect competition: the economics of Russian roulette', *De Economist*, **144**(2), 223–58.
Bröcker, J. (1994), 'Die Lehren der neuen Wachstumstheorie für die Raumentwicklung und die Regionalpolitik', in U. Blien, H. Herrmann and M. Koller (eds), *Regionalentwicklung und regionale Arbeitsmarktpolitik, Konzepte zur Lösung regionaler Arbeitsmarktprobleme?*, Beiträge zur Arbeitsmarkt- und Berufsforschung Nr. 184, Nuremberg: Landesarbeitsamt Nordbayern.
Currie, M. and I. Kubin (1995), 'Non-linearities and partial analysis', *Economic Letters*, **49**, 27–31.
Currie, M. and I. Kubin (2006), 'Chaos in the core–periphery model', *Journal of Economic Behavior and Organization*, **60**, 252–75.
Day, R.H. (1982), 'Irregular growth cycles', *American Economic Review*, **72**, 406–14.
Dixit, A.K. and J.E. Stiglitz (1977), 'Monopolistic competition and optimum product diversity', *American Economic Review*, **67**, 297–308.
Döring, T. and J. Schnellenbach (2006), 'What do we know about geographical knowledge spillovers and regional growth? A survey of the literature', *Regional Studies*, **40**, 375–95.
Fujita, M., P.R. Krugman and A.J. Venables (1999), *The Spatial Economy: Cities, Regions and International Trade*, Cambridge, MA: MIT Press.
Gomes, O. (2007), 'The dynamics of growth and migrations with congestion externalities', *Economic Bulletin*, **15**, 1–8.
Grossman, G.M. and E. Helpman (1991), *Innovation and Growth in the Global Economy*, Cambridge, MA: MIT Press.
Hicks, J.R. (1939), *Value and Capital: An Inquiry into some Fundamental Principles of Economic Theory*, Oxford: Clarendon Press.
Hirschman, A.O. (1958), *The Strategy of Economic Development*, New Haven, CT: Yale University Press.
Johansson, B. and J.M. Quigley (2004), 'Agglomeration and networks in spatial economics', *Papers in Regional Science*, **83**, 165–76.
Keeble, D. and F. Wilkinson (2000), *High-Technology Clusters, Networking and Collective Learning in Europe*, Aldershot: Ashgate.
Krugman, P.R. (1991a), *Geography and Trade*, Cambridge, MA: MIT Press.
Krugman, P.R. (1991b), 'Increasing returns and economic geography', *Journal of Political Economy*, **99**, 483–99.
Lim, U. (2007), 'Knowledge externalities, spatial dependence, and metropolitan economic growth in the United States', *Environment and Planning A*, **39**, 771–88.
Lin, S.A. (1976), *Theory and Measurement of Economic Externalities*, New York: Academic Press.
Lipsey, R.G. and K. Lancaster (1957), 'The general theory of second best', *Review of Economic Studies*, **24**, 11–32.
Maier, G. (2001), 'History, spatial structure, and regional growth: lessons for policy making', in B. Johansson, Ch. Karlsson and R.R. Stough (eds), *Theories of Endogenous Regional Growth: Lessons for Regional Policy*, Berlin: Springer, pp. 111–34.
Maier, G. and S. Sedlacek (eds) (2005), *Spillovers and Innovation: Space, Environment, and the Economy*, Vienna and New York: Springer.
Malmberg, A. and P. Maskell (2002), 'The elusive concept of localization economies: towards a knowledge-based theory of spatial clustering', *Environment and Planning A*, **34**, 429–49.
Malmberg, A. and P. Maskell (2006), 'Localized learning revisited', *Growth and Change*, **37**, 1–18.
Marshall, A. (1920), *Principles of Economics*, 8th edn, London: Macmillan.
McCann, P. and D. Shefer (2004), 'Location, agglomeration and infrastructure', *Papers in Regional Science*, **83**, 177–96.
Mills, E.S. (1972), 'An aggregative model of resource allocation in a metropolitan area', in M. Edel and J. Rothenburg (eds), *Readings in Urban Economics*, New York: Macmillan, pp. 112–23.
Mishan, E.J. (1971), 'The postwar literature on externalities: an interpretive essay', *Journal of Economic Literature*, **9**, 1–28.
Myrdal, G. (1957), *Economic Theory and Underdeveloped Regions*, London: Duckworth.
Peitgen, H-O., H. Jürgens and D. Saupe (1998), *Chaos, Bausteine der Ordnung*, Reinbeck bei Hamburg: Rowohlt.
Ottaviano, G.I.P. and D. Puga (1997), 'Agglomeration in the global economy: a survey of the "new economic geography" ', Centre for Economic Performance, Discussion Paper No. 356.
Polya, G. (1931), 'Sur quelques points de la théorie de probabilités', *Annales Institute H. Poincaré*, **1**, 117–61.
Puu, T. (1997), *Mathematical Location and Land Use Theory: An Introduction*, Berlin: Springer.
Puu, T. (2000), *Attractors, Bifurcations, and Chaos*, Berlin: Springer.
Romer, P.M. (1986), 'Increasing returns and long run growth', *Journal of Political Economy*, **94**, 1002–37.
Romer, P.M. (1990), 'Endogenous technological change', *Journal of Political Economy*, **98**, S71–S102.
Schumpeter, J.H. (1954), *History of Economic Analysis*, London: George Allen & Unwin.
Solow, R.M. (1956), 'A contribution to the theory of economic growth', *Quarterly Journal of Economics*, **70**, 65–94.

Starrett, D. (1978), 'Market allocations of location choice in a model with free mobility', *Journal of Economic Theory*, **17**, 21–37.
Swan, T.W. (1956), 'Economic growth and capital accumulation', *Economic Record*, **32**, 334–61.
Venables, A.J. (2006), 'Economic geography', in B.R. Weingast and D. Wittman (eds), *The Oxford Handbook of Political Economy*, Oxford: Oxford University Press, pp. 739–54.
Yousefi, S., Y. Maistrenko and S. Popovych (2000), 'Complex dynamics in a simple model of interdependent open economies', *Discrete Dynamics in Nature and Society*, **5**, 161–77.

4 Regional growth and trade in the new economic geography and other recent theories

Kieran P. Donaghy

4.1 Introduction

Trade, international or interregional, is essentially the exchange of goods and services over space. By definition, then, it involves transportation and, hence, some transaction costs. Perhaps since the time of the Phoenicians (circa 1200 BC), if not before, trade has been viewed as an engine of expansion of national and regional economies, and many growth theories and growth policies have been predicated on the assumption that growth is export-led. In this time of increasing global economic integration, it is perhaps an article of faith that, as Armstrong and Taylor (2000) put it: 'Regions, like nations, must actively trade if they are to be prosperous' (p. 119). But, even as regional economists have looked to international trade theory for accounts of trade-induced growth, growth per se has not been the featured explanandum of trade theory. Rather, the principal intent of trade theory has been to explain patterns of trade and why countries or regions tend to specialize in certain trade-oriented industries. Where interest has been shown in growth by trade theorists, it has more often than not been in how growth shocks – in the forms of technological progress or increased availability of a factor – have affected trade patterns (Gandolfo, 1998). Of course, it is not difficult to reason further from changes in trade patterns to inferences about what the implications may be for a regional economy, as such insightful if iconoclastic theorists as Hirschman (1958) have been ready and willing to do.

From the writings of Smith (1776 [1976]) and Ricardo (1817 [1951]) onward through those of Heckscher (1919 [1949]), Ohlin (1933), Samuelson (1948) and scholars working in the neoclassical tradition into the late 1970s, explanations of trade patterns and specializations were given in terms of comparative advantage that, for the most part, lay in differences in either technology or resource endowments. In addition to the assumption of constant returns to scale in production and competitive markets, these explanations shared the further assumptions that the goods and/or services exchanged between countries or regions were products of different industries and that the economies of the trading countries and regions were dissimilar in structure. A stubborn fact of modern trade, however, is that many of the goods and services exchanged are similar in nature and are exchanged between countries or regions of countries with developed economies (Grubel and Lloyd, 1973). Framers of so-called 'new theories of trade' have sought to account for this rise in 'intra-industry trade'. Some 'neo-orthodox' theorists, such as Falvey and Kierzkowski (1987) and Findlay (1995), have done so by modifying the assumptions of orthodox theories of comparative advantage, while others, such as Krugman (1979, 1980, 1981), Lancaster (1980), Ethier (1982) and Venables (1996), have done so by introducing to their models increasing-returns-to-scale technologies and monopolistic competition.

Modeling modifications along the latter lines have also been employed, with assumptions of costly trade and mobile factors, by proponents of the so-called 'new economic geography' (NEG) to develop micro-behaviorally based explanations of patterns of economic agglomeration in space. These models, which centrally involve trade in finished goods and (in some cases) intermediates, have for the most part been static in orientation, being intended to convey a sense of long-run equilibrium settlement patterns. Fujita and Thisse (2002), however, have argued that, to the extent agglomeration is coextensive with growth, one may view these models as providing explanations of economic growth through trade.

A second generation of publications in the NEG tradition – by Walz (1996), Martin and Ottaviano (1999, 2001), Baldwin and Forslid (2000), Fujita and Thisse (2002) and Yamamoto (2003) – has explicitly modeled a dynamic version of the NEG story by integrating features of the endogenous growth model of Grossman and Helpman (1991) with the Spence–Dixit–Stiglitz machinery.[1] These models have been useful in formally relating trade to growth in regional economies and in indicating how different phases of development can emerge, even though the trade–growth relationship has not been the central plot.

Two implications of the NEG story are that the reduction of costs of transportation and communication leads to greater agglomeration (through a 'home market effect') and that patterns of agglomeration can continue to evolve as costs of transportation and communication change. Krugman and Venables (1995) have provided a model that demonstrates how globalization, as greater integration of the global economy through trade, can give rise to regional agglomerations of activities that promote inequality. But, as Feenstra (1998) has observed, globalization has been characterized not only by integration through trade but also by the disintegration of production of complete goods or the rise of outsourcing. So a further challenge to trade theorists, spatial economists and economic geographers is to explain the fragmentation of production activities over space and to indicate what the implications of the burgeoning trade in semi-finished goods as parts and tasks – as Grossman and Rossi-Hansberg (2006a, 2006b) put it – are for regional economies. Meeting this challenge entails both development in models and empirics. Implications of the growing intra-product trade have been analyzed by theorists of both the neo-orthodox school and by new economic geographers with models based on different assumptions.

The purpose of this chapter is to provide a selective survey of different aspects of the relationship between trade and regional growth that existing theories of trade, agglomeration and fragmentation can help us to understand, and to indicate where the frontiers of research on this subject lie. The plan of the chapter is as follows. The next section will provide a discussion of models explicitly relating growth and trade. The models discussed are taken from the literatures on the 'new (or endogenous) growth theory' and the new economic geography. The subsequent section will take up recent efforts to come to terms with growth and the trade in parts and tasks, or fragmentation, that characterizes globalization. In each section we shall discuss several representative papers in detail and several others in summary fashion. This style of treatment of the material is dictated by the limitations of space and the nature of the exercise. The chapter will conclude with a brief sketch of a suggested research agenda.

4.2 Regional growth and trade

In this section we review a set of models that are intrinsically dynamic and which emphasize the relationship between growth in regional (or what could be construed as 'regional')

economies and trade. The models hail from the literatures of the 'new growth theory' associated with Grossman and Helpman (1991) and Romer (1990) and the new economic geography associated with Krugman (1991) and Venables (1996), *inter alia*.[2]

Rivera-Batiz and Romer (1991)

In their 1991 paper on economic integration and endogenous growth, Luis Rivera-Batiz and Paul Romer (henceforth, RBR) embark from the observation that: 'Many economists believe that increased economic integration between the developed economies of the world has tended to increase the long-run rate of growth' (p. 531). The closer integration thought to promote a higher growth rate can be achieved either by increasing trade in goods or by increasing the flow of ideas. In the models RBR consider, a research and development (R&D) sector with increasing returns to scale (IRS) is the source of growth. An important distinction, operative in the endogenous growth literature and upon which their analysis rests, is that between a 'one-shot gain', which is a 'level effect', and a permanent change in the growth rate, which is a 'growth effect'. RBR note that conventional attempts to quantify effects of integration by scholars using the neoclassical growth model suggest small gains. Their hunch is that estimates calculated in the context of an endogenous growth model would be larger. Noting that the growth effects of trade restrictions have been demonstrated to be complicated, they narrow the focus of their paper and do not consider the more general case of trade between countries with different endowments and technologies.

Following Romer's (1990) specification of production technology, manufacturing output is a function of human capital, H, labor, L, and a set of capital goods, $x(i)$, indexed by i, a continuous variable. Technical progress is represented by the invention of new types of capital goods. There are two types of production activities: production of consumer goods and production of capital goods. Research and development activity creates designs for new types of capital goods. Both types of manufacturing activities employ the same technology, having the form:

$$Y(H,L,x(\cdot)) = H^{\alpha}L^{\beta}\int_0^A x(i)^{1-\alpha-\beta}di, \tag{4.1}$$

in which Y denotes output and A is the index of the most recently invented good. Given the definition of A, $x(i) = 0$ for all $i > A$.

In view of the common production function shared by consumption and capital goods, the relative prices of all goods are fixed by technology and are set to unity, implying that the aggregate capital stock, $K = \int_0^A x(i)di$, and aggregate output, Y, are well defined. The manufacturing decision is separated from the monopoly pricing decision of patent holders. The division of inputs between sectors can be described by the following adding-up constraint, $Y = C + K$.

The institutional arrangements are as follows. There are many firms that rent capital goods from patent holders, hire unskilled labor, and employ skilled human capital to produce manufactured goods. Each firm can produce consumption goods for sale to consumers and produce one capital good on contract for the holder of the good's patent. All manufacturing firms are price-takers, earning zero profit, and manufacturing output is the numeraire. The firm holding the patent on good j bids out the production of capital goods to a manufacturer and purchases physical units of the good for the competitive price (1.0).

The patent holder then rents out the units to all manufacturing firms at the profit-maximizing monopoly rental rate. Patents are tradable assets with prices, P_A, equal to the present value of the stream of monopoly rents payable to the patent holder, minus the cost of the machine embodying the patented technology.

RBR consider two specifications of R&D. In the first, the stock of human capital, H, and engineering knowledge, A, are the only inputs influencing the output of designs,

$$\dot{A} = \delta HA. \tag{4.2}$$

Given the factor-intensity difference between manufacturing and R&D, the model with the knowledge-driven specification of R&D must be analyzed with a two-sector framework. In the second specification, the technology for R&D uses the same inputs as the manufacturing technology in the same proportions:

$$\dot{A} = BH^{\alpha}L^{\beta}\int_0^A x(i)^{1-\alpha-\beta}di. \tag{4.3}$$

Whereas, in this case, human capital, unskilled labor and capital goods are productive in research, knowledge, A, as such is not. RBR refer to this as the 'lab-equipment' specification of R&D. Total output in the second case can be written as:

$$C + \dot{K} + \dot{A}/B = H^{\alpha}L^{\beta}\int_0^A x(i)^{1-\alpha-\beta}di. \tag{4.4}$$

In the knowledge-driven model, output of designs is homogeneous of degree two (HD2) ruling out marginal-product compensation of both A and H. RBR assume that A receives no compensation; hence, designers of new goods may exploit ideas in existing designs. In the lab-equipment model, output of design is HD1. In both models, the stocks of physical capital and knowledge evolve, whereas those of labor and human capital are given.[3] In the knowledge-driven model, it is as if research were done by independent researchers who use their human capital to produce designs for sale. In the lab-equipment model, it is as if R&D were undertaken by separate firms that hire inputs, produce patentable designs for sale.

RBR solve for the balanced growth rate as the rate that equates the interest rate implied by the equilibrium in the production sector, $r_{technology}$, which varies by model, and the interest rate implied by the representative consumer's first-order conditions for intertemporal optimization under the assumption of Ramsey preferences with constantly elastic utility, $r_{preferences}$. In the case of knowledge-driven growth, the balanced growth rate is given by

$$g = (\delta H - \wedge\rho)/(\wedge\sigma + 1), \tag{4.5}$$

where $\wedge = \alpha(\alpha+\beta)^{-1}(1-\alpha-\beta)^{-1}$, and σ is the consumer's intertemporal elasticity of substitution and ρ is the consumer's temporal discount rate. In the case of growth driven by the trade in lab-equipment, the balanced growth rate is:

$$g = (\Gamma H^{\alpha}L^{\beta} - \rho)/\sigma, \tag{4.6}$$

where $\Gamma = B^{\alpha+\beta}(\alpha+\beta)^{\alpha+\beta}(1-\alpha-\beta)^{2-\alpha-\beta}$. Equations (4.5) and (4.6) suggest that scale effects are the only lasting source of gains from trade and economic integration.

RBR conduct thought experiments to address three questions:

1. Can free trade in goods between countries induce the same increase in the balanced growth rate as complete integration into a single economy?
2. If not, can the free movement of goods combined with the free movement of ideas reproduce the rate of growth under full integration?
3. What is the underlying explanation for the dependence of the growth rate on the extent of the market?

To answer these questions, RBR start their thought experiments with the assumption that two isolated economies are growing at the balanced growth rate. They first allow for trade in goods, but restrict the flow of ideas. The answer they find to their first question is that, under the assumptions made, trade in goods has no effect on the long-run rate of growth. In the second experiment, they calculate the additional effect of opening communications networks and permitting flows of ideas. The answer they find to their second question is that allowing the flow of ideas results in a permanently higher growth rate. In the third experiment, RBR consider the effects of opening trade in goods under the lab-equipment specification. In this case trade in goods alone causes the same permanent increase in the rate of growth as complete integration, and the flow of ideas has no additional effect.

In the first experiment (with flows of goods but not ideas in the knowledge-driven model) the only trades that take place are exchanges of capital goods produced in one country for capital goods produced in the other. Free trade in goods does not affect the split of human capital between manufacturing and research. Hence, it does not change the balanced rate of growth or the interest rate. But free trade in goods can affect the level of output (and therefore welfare), however.

In the second experiment (with flows of information in the knowledge-driven model) greater flows of ideas permanently increase the rate of growth. Increasing the flow of ideas has the effect of doubling the productivity of research in each country. Flows of ideas and goods together have the same effect on the growth rate as complete integration does. Whereas complete integration would permit permanent migration, migration is not necessary to achieve productive efficiency.

In the third experiment (with flows of goods in the lab-equipment model), opening trade in goods would cause the same kind of increase in profit earned at each date by the holder of a patent if the interest rate remained constant. In this case, $r_{technology}$ increases by a factor of $2^{\alpha+\beta}$. The lab-equipment model shows that local knowledge spillovers are unnecessary to speed up growth. Also in the lab-equipment model, IRS comes about from the fact that the fixed cost that must be incurred to design a new good is incurred only once when there is integration. Both the knowledge-driven and the lab-equipment models exhibit IRS in the production of new designs as a function of the stocks of the basic inputs.

The key finding of RBR's study is that:

> In a model of endogenous growth, if economic integration lets two economies exploit increasing returns to scale in the equation that represents the engine of growth, integration will raise the long-run rate of growth purely because it increases the extent of the market. . . . [T]his integration could take the form of trade in goods, flows of ideas, or both. (p. 550)

Walz (1996)

In his 1996 paper on transport costs, intermediate goods and localized growth, Uwe Walz presents a dynamic, two-region general equilibrium model in which interregional production and trade patterns are determined. The question motivating his analysis is: 'What is the impact of further integration on regional development and trade specialization patterns?' By employing an endogenous growth approach in a regional model with factor mobility and transport costs, Walz attempts both to narrow the gap in the literature concerning his question and to investigate the process of adjustment towards a long-run equilibrium. As in the previous paper discussed, and unlike in Richardson (1973), growth in Walz's model does not disappear in the long run. Following Grossman and Helpman (1991) and Romer (1990), growth stems from permanent product innovation in the intermediate-goods sector, leading to an ever-growing variety of intermediate goods and services. The growing number of inputs results in higher productivity of final goods production. Intermediates are not traded between regions. Concentration results from interaction between transport and fixed costs. Skilled workers are mobile but unskilled workers are not. The model formalizes Pred's (1966) cumulative causation story and the role of Hirschman's (1958) backward and forward linkages in regional specialization. Walz's analysis suggests that linkages between intermediate- and final-goods producers can create a core–periphery pattern and that a home market effect will tend to be observed.

The essential details of the model are as follows. There are two regions between which transport of intermediate goods is costly. There is an industrial good, Y, and a traditional good, Z. Exchange of final goods is costless. Each region has an endowment of a stock of an immobile factor, $\overline{L}$, which may be construed as either land or unskilled workers. There is also a stock of mobile skilled workers, $\overline{M}$. The location of final demand is of no consequence; mobile workers migrate to the region in which the highest wage is paid. In addition to markets for land and labor, there is also a market for capital. Private households use savings to purchase stocks. Innovating firms finance research outlays through a stock market. All consumers have the same intertemporal utility function at time t:

$$U_t = \int_t^{\infty} e^{-\rho(\Gamma - t)}(\nu \ln C_y(\Gamma) + (1 - \nu)\ln C_z(\Gamma))d\Gamma,\ 0 < \nu < 1, \tag{4.7}$$

in which $C_Y(\Gamma)$ *and* $C_z(\Gamma)$ represent consumption levels of the final goods at time Γ, and ρ is the subjective discount rate. Consumers maximize (4.7) subject to an intertemporal budget constraint. Because of free trading, prices and interest rates are the same in both regions. The value of the household portfolio is denoted by $V(t)$. In contrast with many of the new economic geographers, who stress immobile demand as a force of dispersion, Walz emphasizes the importance of supply-side factors.

Optimizing behavior by consumers leads as a result to the standard Keynes–Ramsey rule:

$$\dot{E}/E = r - \rho, \tag{4.8}$$

according to which, expenditure will be increasing or decreasing over time as the interest rate exceeds or is less than the rate of time preference. Normalizing so that $E = 1$ results in steady expenditure, hence, equality of the interest and discount rates, $r = \rho$. Static demand functions can then be derived as $\nu = P_y C_y$, and $(1 - \nu) = P_z C_z$.

Turning to the production side of the market, the traditional good, Z, is produced by a Cobb–Douglas technology:

$$Z^i = (L_Z^i)^{\delta}(M_Z^i)^{1-\delta}, \tag{4.9}$$

in which L_Z^i and M_Z^i denote land and labor used in producing Z in region i. In addition to these primary inputs, production of the industrial good, Y, also involves the available set of intermediate inputs, employed according to Ethier's (1982) nested technology.

$$Y^i = (M_Y^i)^{\alpha}(L_Y^i)^{\beta}\left[\int_0^n s^i(\iota)^{\gamma} d\iota\right]^{(1-\alpha-\beta)/\gamma}, \text{ with } 0 < \alpha,\beta,\gamma, \tag{4.10}$$

in which n is the number of known intermediate goods and $s^i(\iota)$ denotes the amount of the ι'th intermediate good used in the production of Y. With producers taking n as given, there is perfect competition in the final-goods sector. The intermediate-goods and R&D sectors use only the mobile factor in production, hence $x(\iota) = M_x(\iota)$.

Because of patents, there is only one producer of each intermediate good. Each intermediate good producer in region i faces a demand function in region j which follows from profit maximization by final-goods producers:

$$s^{i,j} = \frac{(p_x^{i,j})^{\varepsilon}}{\int_0^n p_x^j(\iota)^{1-\varepsilon} d\iota}(1-\alpha-\beta)p_Y Y^j, \text{ with } i, j = \text{A, B.} \tag{4.11}$$

If production of final and intermediate goods occurs in different regions, transport costs arise. Walz assumes them to be of the iceberg type, hence, for every one unit of intermediate good shipped, only k units $(0 < k < 1)$ arrive. On this formulation, transport costs are paid by the region of origin and the elasticity of demand is unaffected. Producer (free on board – FOB) and user (cost insurance and freight – CIF) prices are then as follows:

$$p_x^{i,j} = \begin{cases} q_x^i, & \text{if } i = j, \\ q_x^i/k & \text{if } i \neq j. \end{cases} \tag{4.12}$$

The output of region i producers sold in j is $x^{i,j} = s^{i,j}/k$, *if* $i \neq j$ and $x^{i,j} = s^{i,j}$, if $i = j$. Pulling this all together, the total production of a producer in region i can be expressed by:

$$x^i = (q_x^i)^{-\varepsilon}(1-\alpha-\beta)p_Y\left[\frac{Y^j k^{\varepsilon-1}}{\int_0^n p_x^j(\iota)^{1-\varepsilon} d\iota} + \frac{Y^j}{\int_0^n p_x^i(\iota)^{1-\varepsilon} d\iota}\right]. \tag{4.13}$$

Maximized profit flows for intermediate-goods producers in region i are:

$$G^i = (q_x^i - c_x^i)x^i, \tag{4.14}$$

where $c_x^i = w^i$ denotes production costs of region i intermediate-goods producers, w^i being the wage rate of mobile labor in the region. (The immobile factor is paid the same in each region.) From (4.13) and (4.14), Walz obtains for the optimal price:

$$q_x^i = w/\gamma. \tag{4.15}$$

Producer prices for all intermediate-good varieties, then, are the same irrespective of the region of consumption.

Forward-looking firms invest in R&D and are compensated for future profits. Modeling of the R&D sector follows Grossman and Helpman (1991) and Romer (1990). If an entrepreneur in region i employs M_n^i units of labor at any point in time, he or she is able to invent new intermediate goods, n^i, at the rate:

$$\dot{n} = \frac{M_n^i}{a_n} K^i \tag{4.16}$$

per unit of time. In (4.16), K^i is the stock of knowledge in the respective region and represents the spillovers of knowledge from the innovation process. Assuming international spillovers, ideas spread evenly and between regions. In this model, 'knowledge spillovers contribute to the permanent growth process but not to the geographical concentration of industrial activity' (p. 678).

Entrepreneurs invest in R&D if the gains from a new product meet or exceed costs. Given free market entry, gains and costs will be offsetting if the rate of innovation is positive. The growth rate of the economy, g, is defined in terms of the rate of innovation, $g = \dot{n} / n$ = rate of innovation.

Households will invest in innovative firms if profits per share, corrected for expected losses or gains, equal the subjective discount rate: $G^i/v^i + \dot{v}^i/v^i = \rho$. This arbitrage condition characterizes a capital-market equilibrium. The goods market clearing conditions are $v = P_Y(Y^A + Y^B)$, and $(1 - v) = P_Z(Z^A + Z^B)$. The economy-wide market clearing condition for the mobile factor is $\overline{M} = M_X + M_Y + M_Z + M_N$, and the regional factor-market clearing condition for the immobile factor is $\overline{L}^i = L_Y^i + L_Z^i$.

A long-run, steady-state equilibrium is characterized by a constant intersectoral and interregional factor allocation. Walz defines $\mu = n^A/n^B$ to be the number of new intermediate goods produced in region A vis-à-vis region B. There are two types of long-run equilibria: one in which innovation is concentrated in one region and one in which it takes place in both regions. In the latter case, a steady state occurs only if $g^A = g^B = g = \dot{n}/n$ and μ is constant. In a steady-state equilibrium, innovation is profitable in both regions only if profits are equal in the two regions.

Walz further defines regional production shares of the two final goods as $s_Z^i = Z^i/(Z^A + Z^B)$ and $s_Y^i = Y^i/(Y^A + Y^B)$ and the relative economy-wide demand for the immobile factor in the production of Y and Z as $b = \beta v/(\delta(1 - v))$. Then from the short-run solution of the model, Walz finds that: 'In a world with positive transportation costs for intermediate goods, the production of final goods is relatively concentrated in the region where more intermediate goods are assembled in order to minimize transport costs' (p. 682).

The following relationship between regional profits from intermediate goods production is key to determining the paper's most important results:

$$G^A - G^B = \Omega[s_Y^A \psi^{-1} - s_Y^B], \tag{4.17}$$

where $\psi \equiv \dfrac{\mu + k^{\varepsilon-1}}{\mu k^{\varepsilon-1} + 1}$, and $\Omega = \dfrac{(1 - \alpha - \beta)(1 - \gamma)v(1 - k^{\varepsilon-1})}{n^B + n^A k^{\varepsilon-1}}$.

The relationship (4.17) reveals two basic forces at work:

1. The demand (or market-size) effect is reflected by the market share of the region in production of *Y*. Says Walz: 'A larger market share of the home region in industrial final-goods production provides a larger local market to which intermediate-goods producers can sell their products' (p. 682) with transportation cost savings add-ons.
2. The competition effect implies the opposite outcome with an interior solution. This effect results from the fact that because of transportation costs, competition is more fierce in the region with a larger number of locally produced intermediate goods. An increase in μ, increases competition in A.

In an interior (core–periphery) solution, the competition (demand) effect dominates.

Walz considers three different cases. In the first, large factor endowment differentials lead to a unique equilibrium with a core–periphery pattern. In the second and third, where the regions are equally sized, there can be either an interior or a core–periphery solution. In this model, sectoral shocks and policies will affect regions asymmetrically, leading Walz to observe: 'A policy designed to bring regions closer together actually increases the gap between their production and growth structures' (p. 690).

Owing to space limitations, the discussion of the remaining papers in this section will be less detailed. Other dynamic models in the NEG tradition not discussed here, but discussed in Berliant and Wang (2004) are Chapter 11 of Fujita and Thisse (2002) and Baldwin and Forslid (2000).[4]

Martin and Ottaviano (1999, 2001)

In their 1999 paper on industry location in a model of endogenous growth, Philippe Martin and Gianmarco Ottaviano also sought to integrate the general thrust of the NEG with the new growth theory to promote an understanding of the relationship between location and growth where regions persist. Whereas Walz (1996) considered aggregate returns to scale at the local level and migration, his focus on IRS was at the aggregate level, rather than the firm level, which is closer to the concern of the NEG. Martin and Ottaviano (henceforth MO) believe that:

> the process of creation of new firms and the process of location should be thought of as joint processes. When the external effects which are at the source of endogenous growth are local in nature, because they involve localized interactions between economic agents, then the location of firms and of R&D activities will effect the process of technological change. Technological change, when internalized in the creation of new goods and new firms, will in turn have an impact on the extent and direction of foreign direct investment, and more generally capital flows. (1999, p. 282)

In their analysis, MO find that the introduction of explicit dynamics in a locational model changes some of the results found in the NEG literature. They examine how growth affects the location decisions of firms and hence how it effects geography and the dynamics of spatial distribution of economics activities.

In the model they consider, firms can choose to locate between two trading locations (North and South). The two locations are identical except for initial levels of non-labor wealth. The North is wealthier. Each firm requires a new idea, which is created through

R&D. So growth comes about by expansion of product variety. Their location framework differs from the NEG in that cumulative causation mechanisms, such as migration or vertical linkages, are excluded. They analyze the relationship between location and growth in two cases. In the first, spillovers of R&D are global and reduce costs of future R&D in both locations. Economic geography in this case does not affect growth, but costs of R&D and the discount rate do affect income differentials. In the second case, R&D spillovers are local, and R&D costs are lowest where there is the highest number of firms producing differentiated products. In this case all R&D activities agglomerate in the North, where firms are more numerous and the growth rate is higher, the more concentrated the industry.

MO find that when spillovers are global, economic geography does not influence the growth rate. However, high growth rates are associated with capital flows to the South, because the factors that increase growth rates also decrease the differential in income between the North and South. The creation of new firms is the driving force behind capital flows. When spillovers are local, spatial concentration of activities is beneficial to growth, implying that a decrease in transportation costs favors the rate of innovation and growth. Also, when spillovers are local, industrial concentration brings a previously unanalyzed trade-off between aggregate growth and regional equity. On the one hand, an increase in the industrial concentration in the location where R&D is performed increases growth, which is an effect not internalized in the location choice of firms. On the other hand, the welfare cost of transportation between the two locations is minimized when industry is split evenly between the two locations. This finding implies that the net result of a decrease in both transportation costs and R&D costs, leading to spatial concentration, can be a welfare gain. MO also show that the South can gain from more concentration in the North if growth benefits are large enough. If the R&D sector also uses the differentiated goods, then an increase in growth will increase the market size of the innovation location and lead to industrial migration to that location.

Whereas Martin and Ottaviano (1999) did not consider mechanisms of cumulative causation, in their 2001 paper on 'Growth and agglomeration', the authors construct a model in which aggregate growth and spatial agglomeration are mutually reinforcing processes. Innovation-led growth is conducive to spatial agglomeration, which in turn lowers costs of further innovation and promotes a higher rate of growth. If innovative industries use goods from monopolistically competitive industries as inputs, the suppliers will be drawn to locations where innovation occurs, creating a forward linkage. The presence of suppliers at innovation locations reduces transaction costs and the cost of innovation in general, thereby increasing incentives to innovate and creating backward linkages to suppliers. MO note that agglomeration occurs in Krugman (1991) because, when transportation costs are low enough, and an IRS sector uses a specific input, there is mobility between locations. Agglomeration occurs in Venables (1996) because there is intersectoral mobility in the presence of intra-sectoral linkages in the sector with increasing returns. MO show a third way that agglomeration can occur: by introducing endogenous growth along lines of Grossman and Helpman (1991) and Romer (1990). To examine cleanly the effect of endogenous growth on agglomeration, MO do not allow either factor mobility or intra-sectoral vertical linkages in the IRS sector.

In the static, first-generation NEG models, which consider only the spatial distribution of a fixed stock of resources, there is only movement from the periphery to the core. In

Martin and Ottaviano (2001), because new firms are continuously created in the core, relocation dynamics are richer and more realistic. As in Krugman's models, there is a spatial divergence of income levels. But MO also show that the more spatially agglomerated an economy is, the faster it grows in aggregate terms. In their model the geography of economic activities matters for growth even in the absence of local technological spillovers – trade affects growth through geography.

Starting from a situation with no growth or agglomeration, MO show that when the aggregate economy starts growing, the only steady-state outcome is that in which one of the two regions gets all of the innovative activity and most of the industrial production. Their analysis uncovers novel location dynamics. While agglomeration takes place in the core, some firms will prefer to locate and produce in the periphery (where competition is less strong) as new activity is generated in the core.

MO have shown that the same factors that spur growth also spur agglomeration and that the cumulative process identified reinforces the effect that a change of one factor has on both growth and agglomeration. 'In particular . . . a decrease in transaction costs between regions of an economy encourages both agglomeration and growth of activities for the whole economy: the growth effect goes through the impact on geography and the agglomeration effect goes through the impact on growth' (2001, p. 967).

Yamamoto (2003)

We briefly note a more recent paper by Kazuhiro Yamamoto that is closely related to Martin and Ottaviano (2001). Yamamoto (2003) develops a Romer-type endogenous growth model with two countries in which production of a homogeneous manufactured good requires differentiated intermediate goods. In the tradition of such models, 'growth' means expansion of the variety of intermediate goods. There is circular causation between growth and agglomeration – growth yields agglomeration, which reinforces more growth, due to indirect vertical linkages between the innovation and intermediate goods sectors. Innovative activities use the manufactured good as an input, while the manufactured good is produced with differentiated intermediate goods.

The model yields two types of international trade patterns of the manufactured good, which are determined by the relationship between transportation costs of the manufactured good and the intermediate goods. In the first pattern, the manufactured good is produced in both countries but no trade of the good occurs. All innovative activity is concentrated in one country, the 'core', and all firms producing intermediate goods locate there as well. All new firms form in the core, but some relocate to the periphery. In the second type of international trade pattern, manufactured good production concentrates in just one country, where intermediate-good production is also fully concentrated. In this case, the two-country economy achieves the maximal growth rate attainable. The first situation arises if the transportation cost of the manufactured good is sufficiently high, whereas the second occurs if this cost is sufficiently low. The relationship between transportation costs for manufactured goods and those for intermediate goods influences decisions regarding plant locations and growth of the economy. Yamamoto claims that the model can capture the post-war process of industrialization that has occurred in East Asia; his model's solution captures the stylized facts of the process by which a growing variety of intermediate goods has given rise to spatial agglomeration. Remarking upon the role that ICT has played in this process, Yamamoto observes that information may be viewed

as an intermediate good. He also notes that, in the process of growth in a spatial economy, governments may have a role to play in coordinating forward expectations of agents.

4.3 Growth and trade in parts and tasks

In 1998 (already a decade ago), Robert Feenstra observed that there has been 'a spectacular integration of the global economy through trade . . . [T]he world is *much* more integrated today than at any time during the past century' (Feenstra, 1998, p. 31). He continued: 'the rising integration has brought with it a disintegration of the production process, in which manufacturing or service activities done abroad are combined with those performed at home' (Feenstra, 1998, p. 31). Indeed, Yi (2003) appeals to this very disintegration of production, a manifestation of what Austrian economists termed 'increasing roundaboutness', as an explanation for the growth in world trade. Kaminski and Ng (2005) provide a rich study of how the disintegration of production in Central and Eastern Europe has helped the CEEC-10 countries to become net exporters of products and parts in production networks and thereby grow their economies.[5]

The disintegration or (more commonly) the fragmentation of production has been described by a number of expressions; among them are slicing up the value chain, outsourcing, vertical specialization and intra-product specialization. Jones and Kierzkowski (1990, 2001, 2005) are widely credited for identifying and giving theoretical structure to this phenomenon, although Yi (2003) suggests that Balassa (1967) and Findlay (1978) were the first to notice it. Hummels and Levinsohn (1993) are credited with being the first to model the phenomenon formally.

The increase in the trade in intermediate goods and tasks is seen to promote economic growth directly and, to the extent that it occurs at the regional level, induce regional growth (Venables, 2006). In some cases, it is also perceived to have deleterious effects at the regional level. Munro et al. (2007) and Polenski and Hewings (2004) write about the 'hollowing-out' of local industry that occurs as value chains are extended and local linkages dwindle. And there is pervasive concern about loss of jobs and depression of wages at lower skill levels, as outsourcing and offshoring is increased.[6]

To investigate the panoply of regional impacts of fragmentation (by all its names), trade theorists are employing a wide range of modeling frameworks. Some, such as Deardorff (2001) and Yi (2003), employ refinements of Ricardian and Heckscher–Ohlin approaches, in which scale economies and imperfect competition are not appealed to. Others, such as Jones and Kierzkowski in their various publications, suggest that economies of scale and scope are to be found in service blocs that integrate production units. Still others, pursuing the NEG agenda, are employing to good effect models of imperfectly competitive firms that enjoy scale economies at the firm level. And some theorists, such as Grossman and Helpman (2005) and Grossman and Rossi-Hansberg (2006a, 2006b) are helping to refine the micro-behavioral foundations of firm- and inter-firm-level analysis. In this section we present a selection of papers that convey a sense of how fragmentation and its effects on national and regional economies can be analyzed.

Deardorff (2001)

Alan Deardorff examines the effects of fragmentation, in simple trade models of the Ricardian and Heckscher–Ohlin sorts of small open economies in a two-country world, on growth, national welfare, patterns of specialization and trade, and on factor prices.

Although focused at the country level, his findings are directly applicable to regional economies.

Deardorff's working definition of fragmentation is 'the splitting of a production process in two or more steps that can be undertaken in different locations but that lead to the same final product' (Deardorff, 2001, p. 122). Deardorff sees fragmentation as 'a manifestation of globalization and technology combined, since in many industries it is only advances in technology that have made the splitting of production processes and the coordination of the resulting parts [of the production process] possible' (p. 122). He observes that fragmentation can occur both within and across countries, implying interregional and international trade in parts and tasks. The focus of his paper, however, is on international fragmentation in the context of competitive markets.[7]

Deardorff initially assumes a standard Ricardian set-up, in which all intermediate goods are tradable. Fragmentation is allowed to occur within the domestic as well as export industries. A country may produce and export intermediate goods only. He demonstrates that fragmentation may give a country a comparative advantage in a good where it had no comparative advantage before, and allow the world to benefit from alternative combinations of factors and goods on an expanded factor-price frontier. What the Ricardian model cannot address is the existence of separate factors of production and the possibility that some countries may gain while others may lose from such a gain.

Turning next to Heckscher–Ohlin (H–O) frameworks, Deardorff observes that fragmentation will not occur in the kinds of equilibria most often considered in the H–O literature.[8] If, in such frameworks, there is factor-price equalization (FPE), and fragmentation is costly, there will be no incentive to fragment production. This state of affairs implies that in order for fragmentation to be interesting there must be different factor prices in the world economy. However, in the event of different factor prices, fragmentation increases the possibility of FPE occurring.

The main conclusions Deardorff draws from his analysis are:

1. If fragmentation does not change the prices of goods, then it must increase the value of output of any country where it occurs and that of the world. That is, increasing fragmentation – and the associated increased trade in parts and tasks – leads to an increased growth rate.
2. If fragmentation does change prices, then fragmentation can lower the welfare of a country by turning its terms of trade against it.
3. Even in a country that gains from fragmentation, it is possible that some factor owners within a country will lose.
4. To the extent that factor prices are not equalized internationally in the absence of fragmentation, fragmentation may be a force toward factor price equalization.

Ethier (2005)

In his 2005 contribution to a special journal issue on fragmentation, Wilfred Ethier views (aspects of) globalization as the *explanans* (instead of the *explanandum*) for some very prominent stylized facts of the current world economy. He perceives three new foci of concern with respect to globalization. These are: first, the reduction in barriers to economic exchange has proceeded to the point where outsourcing is affecting wages of

unskilled workers in countries with developed economies; second, it has become possible to adjust production methods globally in response to change in the economic environment; and, third, the inclusion of new participants in the multilateral trade system raises the prospect of a fundamental reallocation of global production. Ethier addresses globalization and the so-called 'skill premium', unemployment of unskilled workers, technology and the independence of national social policies.

Taking the measure of prevailing trade theoretic frameworks, Ethier observes that empirical regularities concerning the rise of a skill premium do not comport well with the Stopler–Samuelson explanation.[9] His response is to hypothesize that intersectoral differences in technology have little importance for the relationship between trade and wages. Hence, he infers it might be best to abstract from inter-sectoral relations. Instead, he focuses on the intra-sectoral ease of substitution between assets. Ethier adopts a model in which all firms can use the same techniques. There is constant return to scale, which implies the existence of an aggregate production function, and the purpose of trade in the model is to allow international fragmentation of the production process.

Ethier assumes a high degree of substitution between unskilled domestic labor and outsourcing. He stresses that skilled labor and equipment are complements, and assumes there is intermediate substitutability of an equipment and skilled labor aggregate for unskilled labor and outsourcing. In addition, equipment must be provided for from output. Firms choose freely their degree of equipment utilization. In his model neutral exogenous technical change is possible, as is endogenous skill-biased technical change. Ethier considers trade between two countries (regions), of which one can outsource part of the production process and supply equipment and the other can supply outsourcing but not equipment. He employs the following aggregate production function, AF:

$$AF(U, V, E, S) = A[U^{\gamma} + V^{\gamma}]^{\alpha/\gamma}\,[E^{\sigma} + S^{\sigma}]^{(1-\alpha)/\sigma}, \tag{4.18}$$

in which,

A = index of total factor productivity,
U = employment of unskilled labor (out of a total stock U_0),
V = quantity of tasks outsourced,
E = stock of equipment,
S = employment of skilled labor.

Ethier's central hypothesis is that outsourcing substitutes for unskilled labor, thus $1 > \gamma > 0$. A positive value of γ less than one implies that there are diminishing returns to outsourcing. The assumption that skilled labor and equipment are complements implies that $\sigma < 0$. Outsourcing must be paid for with exports, X. Hence:

$$X = \frac{v^*}{g} V, \tag{4.19}$$

in which v^* is the foreign price of V and g is an index of globalization, $1 \geq g \geq 0$. Nationally owned output (income) is:

$$Y = AF - \frac{v^*}{g}V. \tag{4.20}$$

Equipment services require that a portion of output be devoted to investment, I, $I = E/Q$, where the parameter Q represents the technology for providing equipment services. Consumption is determined by residual:

$$C = AF - \frac{E}{Q} - \frac{v^*}{g}V. \tag{4.21}$$

Ethier operationally defines fragmentation, f, by the following index:

$$f \equiv \frac{V^\gamma}{U^\gamma + V^\gamma}, \tag{4.22}$$

to capture the relative importance of outsourcing to the U-V sub-aggregate. He also defines an equipment utilization index:

$$e \equiv \frac{S^\sigma}{E^\sigma + S^\sigma}, \tag{4.23}$$

to measure the relative importance of equipment in the E-S sub-aggregate. Differentiation of these latter two definitions yields the following useful relations:

$$\frac{V}{f}\frac{df}{dV} = \gamma(1-f), \tag{4.24}$$

$$\frac{U}{f}\frac{df}{dU} = -\gamma(1-f), \tag{4.25}$$

$$\frac{E}{e}\frac{de}{dE} = -\sigma(1-e). \tag{4.26}$$

From inspection of (4.24) and (4.25), it is apparent that an increase in V or a decrease in U raises f (fragmentation) if and only if $\gamma > 0$, where V measures outsourcing and f indexes the extent V displaces U. From (4.19) it can be seen that equipment utilization, e, responds positively to an increase in E if and only if $\sigma < 0$ When each input is paid the value of its marginal product, the following input pricing relations apply:

$$\begin{gathered} w = AF_U = \alpha\frac{AF}{U}(1-f),\ v^*/g = AF_V = \frac{\alpha AF}{V}f, \\ 1/Q = AF_E = (1-\alpha)\frac{AF}{E}(1-e),\ s = AF_S = (1-\alpha)\frac{AF}{S}e, \end{gathered} \tag{4.27}$$

where w denotes the wage paid unskilled labor and s that of skilled. The 'skill premium' in wages is, then, $\pi = s/w > 1.0$.

The exogenous arguments of the model are endowments, U_0 and S, trade parameters, v^* and g, and technology parameters, A and Q, and the endogenous variables are Y, V, E, S, π (or w), and U. Given the exogenous variables, the equilibrium conditions determine the values of Y, V, E, S, and a relation between π (or w) and U. It is assumed that the government pursues an active social policy that causes it to choose a particular $\pi - U$ combination from those available.

The input pricing relations enable the skill premium to be rewritten as:

$$\pi = \frac{s}{w} = \frac{1-\alpha}{\alpha}\frac{U}{S}\frac{e}{1-f}, \tag{4.28}$$

from which Ethier's first three propositions directly follow:

1. An increase in equipment utilization, e, results in an increase in the skill premium, π, if and only if $\sigma < 0$. An increase in fragmentation, f, results in an increase in π, if and only if $\gamma > 0$.
2. An increase in employment of skilled labor, S, coincides with an increase in π, if there is a sufficient increase in e or f, and $\sigma < 0$, $\gamma > 0$.
3. Given α, unskilled labor employment, U, and S, changes in technology or trade will affect π only through their effects on e or f.

Other than these, the most important results of Ethier's analysis are as follows:

- More fragmentation is associated with more elastic demand for unskilled labor if and only if outsourcing and unskilled labor are sufficiently substitutable.
- Given complementarity of skilled labor and equipment, if unskilled labor and outsourcing are sufficiently substitutable, the government will be unable to influence significantly the skill premium with policies involving variations in employment.
- An increase in globalization turns the terms of trade in favor of the South, increasing equipment utilization in the South and outsourcing to the South.
- An increase in home total factor productivity leads to an increase in the skill premium in both the North and South and an increase in equipment utilization in both the North and South.
- If unskilled labor and outsourcing are sufficiently substitutable, the actions of policy-makers in each country will tend to frustrate those of policy-makers in the other, whenever they attempt to pursue divergent objectives.
- If skilled labor and equipment are highly complementary, employment fluctuation in the South will have negligible effects on the skill premium in the North, but employment fluctuations in the North will significantly influence the skill premium in the South.
- If skilled labor and equipment are highly complementary and unskilled labor and outsourcing are highly substitutable, fluctuations in home employment will have little effect on the skill premium in either the North or South when fragmentation is extensive.

Fujita and Thisse (2006), Fujita and Gokan (2005)

Masahisa Fujita and Jacques-François Thisse also examine fragmentation, but from the perspective of the NEG, and with a different question in mind: 'Who gains and who loses?' Fujita and Thisse observe that for various entities (for example regions of undeveloped countries), liberalization of trade and capital flows, which coincide with fragmentation, can be detrimental by promoting more inequality. So they set out to demonstrate that globalization need not have such implications for low-income people and countries. The setting they choose to examine is that of cross-border fragmentation,

in which the firms that are fragmenting their production are imperfectly competitive. They argue that by adopting a general equilibrium model of monopolistic competition of the Dixit–Stiglitz sort they can identify 'the feedback mechanisms by which the decision made by some firms to engage in foreign direct investments affects market conditions, which in turn influence other firms' decisions about their spatial organization' (p. 812). Each firm has a headquarters and a production plant. There is both skilled and unskilled labor in the model. Headquarters employ skilled labor and plants use unskilled labor and headquarters services. Each firm is free to outsource production but must bear additional costs for doing so. These additional spatial costs pertain to both communication and trade. In order for low-wage areas to become more accessible and attractive for production, new information and communications technologies need to be developed as trade costs fall. Within this framework, Fujita and Thisse want to study 'how the spatial division of labor changes when communication and trade costs become lower and what the corresponding implications are for the various groups of workers' (p. 813).

Depending on the combinations of trade and communication costs possible, the model of Fujita and Thisse can manifest a range of dynamics in agglomeration processes, among which is a partial deindustrialization of the core region, pervasively observed in regions of developed countries. With respect to welfare effects, they find ambivalent results. As expected, unskilled workers in the periphery are better off and unskilled core workers worse off as more plants move to the periphery. But surprisingly, skilled workers in the core are also hurt by fragmentation (through an increase in their local price index). Hence, '*both* types of workers living in the core are worse off when firms gradually relocate their plants into the periphery' (p. 814). This finding suggests that fragmentation could help narrow the gap between rich and poor countries (regions). Fujita and Thisse's results also suggest that 'the IT revolution might well lead to job creation in the periphery at the expense of the core regions' (p. 815).

In a related paper – written after but published before Fujita and Thisse (2006) – Fujita and Toshitaka Gokan (2005) extend spatial fragmentation to multiple plants for each product. They find that with decreasing communications costs, firms producing goods with low trade costs tend to concentrate their plants in low-wage countries, whereas firms producing goods with high trade costs tend to have multiple plants serving segmented markets.[10]

4.4 A research agenda

In his review of the 'new trade theory', Giancarlo Gandolfo (1998) remarked that it is comforting to know that international trade theory has given us the tools for coping with all types of market form. To his remark we might add that this body of theory has also given us the tools for coping with all types of returns to scale (and scope). The observation that different trade-based explanations of growth and structural change can be fashioned from these tool sets is neither new nor interesting. What is perhaps both new and interesting is a growing sense of urgency to conduct empirical studies of regional growth and trade to sort out the explanatory power of competing accounts. The dearth of such studies has regularly been explained away in terms of the paucity of interregional trade data – see, for example, Armstrong and Taylor (2000) – and especially data on trade in services.[11] Given the pace with which trade of all sorts is growing and the potential changes this development may create in regions of countries with developed and

developing economies, it is clear that we need both more and better data, and more empirical studies. Such studies should employ models that are sufficiently flexible to test for the competitiveness of markets and returns to scale – at the level of the firm, industry or in service blocs – and to determine what difference these attributes make to the power of trade-based explanations of regional growth. We also need empirically based models that can be used to anticipate the evolution of trade patterns and their likely outcomes (including impacts on labor markets and the physical environment), and investigate how policy interventions might ameliorate or reinforce any ill effects. It is hoped that research organizations functioning at various levels of spatial and political aggregation in both the public and private sectors can work together with concerned scholars to begin to address these critical needs.

Notes

1. The reader is referred to Fujita and Mori's (2005) recent survey.
2. For a useful survey of the voluminous literature on endogenous growth and trade, see Long and Wong (1997).
3. This assumption concerning L and H is made to avoid having to solve a non-linear system of differential equations with growth rates that vary over time.
4. See also Baldwin et al. (2003), pp. 155–89.
5. CEEC-10 is an acronym for the ten central and eastern European countries that were former Soviet republics. See also Egger and Egger (2003) for an econometric study of the transformation of Austria post-1990 and McCann (2007) on the literature of production networks. On the broader topic of fragmentation, see Arndt and Kierzkowski (2001) and the special issue of *International Review of Economics and Finance*, 14, 2005.
6. See also Yomogida (2007), Egger and Egger (2007) and Egger and Kreickemeier (2008).
7. In his second endnote, Deardorff remarks that: 'Even with perfect competition, if a country is lumpy – that is, if different equilibrium wages are paid in different regions of the country . . . then fragmentation may occur across regions for much the same reasons as . . . international fragmentation' (p. 135).
8. Deardorff's analysis suggests that only within the context of a specific trade model can the effects on growth of changes in trade patterns be analyzed.
9. According to the Stolper–Samuelson Theorem, the increase in the relative price of a commodity favors (raises the unit real reward of) the factor used most intensively in the production of the commodity. (See Stolper and Samuelson, 1941, and Gandolfo, 1998.)
10. In a complementary paper, which does not pertain directly to growth and trade, Duranton and Puga (2005) model the stylized facts of firm fragmentation, made possible by improvements in transportation and communications technologies, and the corresponding changes in urban structure. They note that conditions promoting separation of production facilities from headquarters and management facilities include same-sector specialization for production sites and abundant business service employment in headquarters sites. Such a separation provides an incentive for cities to shift from sectoral specialization to functional specialization.
11. The research by Munro et al. (2007) is exceptional in this regard.

References

Armstrong, H. and J. Taylor (2000), *Regional Economics and Policy*, Oxford: Blackwell Publishing.

Arndt, S. and H. Kierzkowski (eds) (2001), *Fragmentation: New Production and Trade Patterns in the World Economy*, Oxford: Oxford University Press.

Balassa, B. (1967), *Trade Liberalization among Industrial Countries: Objectives and Alternatives*, New York: McGraw-Hill.

Baldwin, R.E. and R. Forslid (2000), 'The core–periphery model and endogenous growth: stabilising and destabilising growth', *Economica*, **67**, 307–24.

Baldwin, R., R. Forslid, P. Martin, G. Ottaviano and F. Robert-Nicoud (2003), *Economic Geography and Public Policy*, Princeton, NJ: Princeton University Press.

Berliant, M. and P. Wang (2004), 'Dynamic urban models: agglomeration and growth', in R. Capello and P. Nijkamp (eds), *Urban Dynamics and Growth*, Amsterdam: Elsevier, pp. 533–81.

Deardorff, A.V. (2001), 'Fragmentation in simple trade models', *North American Journal of Economics and Finance*, **12**, 121–37.

Duranton, G. and D. Puga (2005), 'From sectoral to functional urban specialization', *Journal of Urban Economics*, **57**, 343–70.
Egger, H. and P. Egger (2003), 'Outsourcing and skill-specific employment in a small economy: Austria after the fall of the Iron Curtain', *Oxford Economic Papers*, **55**, 625–43.
Egger, H. and P. Egger (2007), 'Outsourcing and trade in a spatial world', *Journal of Urban Economics*, **62**, 441–70.
Egger, H. and U. Kreickmeier (2008), 'International fragmentation: boon or bane for domestic employment', *European Economic Review*, **52**, 116–32.
Ethier, W.J. (1982), 'National and international returns to scale in the modern theory of international trade', *American Economic Review*, **72**, 389–405.
Ethier, W.J. (2005), 'Globalization, globalization: trade, technology, and wages', *International Review of Economics and Finance*, **14**, 237–58.
Falvey, R.E. and H. Kierzkowski (1987), 'Product quality, intra-industry trade and (im)perfect competition', in H. Kierzkowski (ed.), *Protection and Competition in International Trade: Essays in Honor of W.M. Corden*, Oxford: Basil Blackwell, pp. 143–61.
Feenstra, R.C. (1998), 'Integration of trade and disintegration of production in the global economy', *Journal of Economic Perspectives*, **12**, 31–50.
Findlay, R. (1978), 'An "Austrian" model of international trade and interest rate equalization', *Journal of Political Economy*, **86**, 989–1008.
Findlay, R. (1995), *Factor Proportions, Trade, and Growth*, Cambridge, MA: MIT Press.
Fujita, M. and T. Gokan (2005), 'On the evolution of the spatial economy with multi-unit, multi-plant firms: the impact of IT development', *Portuguese Economic Journal*, **4**, 73–105.
Fujita, M. and T. Mori (2005), 'Frontiers of the new economic geography', *Papers in Regional Science*, **84**, 377–405.
Fujita, M. and J.F. Thisse (2002), *Economics of Agglomeration: Cities, Industrial Location, and Regional Growth*, Cambridge: Cambridge University Press.
Fujita, M. and J.F. Thisse (2006), 'Globalization and the evolution of the supply chain: who gains and who loses?', *International Economic Review*, **47**, 811–36.
Gandolfo, G. (1998), *International Trade Theory and Policy*, Heidelberg: Springer.
Grossman, G.M. and E. Helpman (1991), *Innovation and Growth in the World Economy*, Cambridge, MA: MIT Press.
Grossman, G.M. and E. Helpman (2005), 'Outsourcing in a global economy', *Review of Economic Studies*, **72**, 135–59.
Grossman, G.M. and E. Rossi-Hansberg (2006a), 'The rise of offshoring: it's not wine for cloth anymore', paper prepared for the symposium sponsored by the Federal Reserve Bank of Kansas City on The New Economic Geography: Effects and Policy Implications, Jackson Hole, Wyoming, 24–26 August.
Grossman, G.M. and E. Rossi-Hansberg (2006b), 'Trading tasks: a simple theory of offshoring', working paper.
Grubel, H.G. and P.J. Lloyd (1973), *Intra-industry Trade*, Basingstoke: Macmillan.
Heckscher, E. (1919), 'The effect of foreign trade on the distribution of income', *Ekonomisk Tidskrift*, **21**, 497–512; reprinted in abridged form (1949) in H.S. Ellis and L.A. Metzler (eds), *Readings in the Theory of International Trade*, Homewood, IL: Irwin, pp. 272–300.
Hirschman, A. (1958), *The Strategy of Economic Development*, New Haven, CT: Yale University Press.
Hummels, D. and J. Levinsohn (1993), 'Product differentiation as a source of comparative advantage', *American Economic Review*, **83**, 445–559.
Jones, R.W. and H. Kierzkowski (1990), 'The role of services in production and international trade: a theoretical framework', in R.W. Jones and A.O. Krueger (eds), *The Political Economy of International Trade*, Oxford: Basil Blackwell, pp. 31–48.
Jones, R.W. and H. Kierzkowski (2001), 'Globalization and the consequences of international fragmentation', in R. Dornbusch, G. Calvo and M. Obstfeld (eds), *Money, Factor Mobility, and Trade*, Cambridge: MIT Press, pp. 365–83.
Jones, R.W. and H. Kierzkowski (2005), 'International trade and agglomeration: an alternative framework', *Journal of Economics*, Supplement 10, 1–16.
Kaminski, B. and F. Ng (2005), 'Production disintegration and integration of Central Europe into global markets', *International Review of Economics and Finance*, **14**, 377–90.
Krugman, P.R. (1979), 'Increasing returns, monopolistic competition, and international trade', *Journal of International Economics*, **9**, 469–79.
Krugman, P. (1980), 'Scale economies, product differentiation, and the pattern of trade', *American Economic Review*, **70**, 950–59.
Krugman, P.R. (1981), 'Intraindustry specialization and the gains from trade', *Journal of Political Economy*, **89**, 959–73.
Krugman, P.R. (1991), 'Increasing returns and economic geography', *Journal of Political Economy*, **99**, 483–99.
Krugman, P.R. and A.J. Venables (1995), 'Globalization and the inequality of nations', *Quarterly Journal of Economics*, **110**, 857–80.

Lancaster, K. (1980), 'Intra-industry trade under perfect monopolistic competition', *Journal of International Economics*, **10**, 151–75.
Long, N.V. and K.-Y. Wong (1997), 'Endogenous growth and international trade: a survey', in B.S. Jensen and K.-Y. Wong (eds), *Dynamics, Economic Growth, and International Trade*, Ann Arbor, MI: University of Michigan Press, pp. 11–74.
Martin, P. and G.I.P. Ottaviano (1999), 'Growing locations: industry location in a model of endogenous growth', *European Economic Review*, **43**, 281–302.
Martin, P. and G.I.P. Ottaviano (2001), 'Growth and agglomeration', *International Economic Review*, **42**, 947–68.
McCann, P. (2007), 'Technology, information and the geography of global and regional trade', in R. Cooper, K. Donaghy and G. Hewings (eds), *Globalization and Regional Economic Modeling*, Berlin: Springer-Verlag, pp. 15–34.
Munro, D.K., G.J.D. Hewings and D. Guo (2007), 'The role of intraindustry trade in interregional trade in the Midwest of the US', in R. Cooper, K. Donaghy and G. Hewings (eds), *Globalization and Regional Economic Modeling*, Berlin: Springer-Verlag, pp. 87–105.
Ohlin, B. (1933), *Interregional and International Trade*, Cambridge, MA: Harvard University Press.
Polenski, K.R. and G.J.D. Hewings (2004), 'Trade and spatial economic interdependence', *Papers in Regional Science*, **83**, 269–89.
Pred, A. (1966), *The Spatial Dynamics of US Urban-Industrial Growth*, Cambridge, MA: MIT Press.
Ricardo, D. (1817), *On the Principles of Political Economy and Taxation*, London: J. Murray; reprinted (1951) in P. Sraffa (ed.), *The Works and Correspondence of David Ricardo*, Vol. 1, Cambridge: Cambridge University Press.
Richardson, H. (1973), *Regional Growth Theory*, London: Macmillan.
Rivera-Batiz, L.A. and P.M. Romer (1991), 'Economic integration and endogenous growth', *Quarterly Journal of Economics*, **106**, 531–55.
Romer, P.M. (1990), 'Endogenous technical change', *Journal of Political Economy*, **98**, S71–S102.
Samuelson, P.A. (1948), 'International trade and the equalization of factor prices', *Economic Journal*, **58**, 163–84.
Smith, A. (1776), *An Inquiry into the Nature and the Causes of the Wealth of Nations*, reprinted (1976), Oxford: Clarendon Press.
Stolper, W.F. and P.A. Samuelson (1941), 'Protection and real wages', *Review of Economic Studies*, **9**, 50–73.
Venables, A. (1996), 'Equilibrium locations of vertically linked industries', *International Economic Review*, **37**, 341–59.
Venables, A. (2006), 'Shifts in economic geography and their causes', paper prepared for the 2006 Jackson Hole Symposium.
Walz, U. (1996), 'Transport costs, intermediate goods, and localized growth', *Regional Science and Urban Economics*, **26**, 671–95.
Yi, K.-M. (2003), 'Can vertical specialization explain the growth of world trade?', *Journal of Political Economy*, **111**, 52–102.
Yamamoto, K. (2003), 'Agglomeration and growth with innovation in the intermediate goods sector', *Regional Science and Urban Economics*, **33**, 335–60.
Yomogida, M. (2007), 'Fragmentation, welfare, and imperfect competition', *Journal of Japanese and International Economies*, **21**, 365–78.

5 Endogenous growth theories: agglomeration benefits and transportation costs

G. Alfredo Minerva and Gianmarco I.P. Ottaviano[1]

5.1 Introduction

The role of infrastructure in global and local economic development can hardly be overstated (World Bank, 1994). In particular, as reported by Calderon and Serven (2004), its role has been stressed along two main dimensions: its effects on economic growth and its effects on income inequality. Along the first dimension, most studies focus on the impact of infrastructure on aggregate output, finding it positive. This is highlighted in a seminal contribution by Aschauer (1989), who finds that the stock of public infrastructure capital is a significant driver of aggregate total factor productivity (TFP). Even though subsequent efforts question Aschauer's quantitative assessment, overall his qualitative insight survives more sophisticated econometric scrutiny (see, for example, Gramlich, 1994; Röller and Waverman, 2001). In particular, Calderon and Serven (2003) identify positive and significant impacts on output of three types of infrastructure (telecommunications, transport and energy) and show that such impacts are significantly higher than those of non-infrastructure capital.

The link between infrastructure and long-run growth is much less explored. Easterly and Rebelo (1993) find that public expenditure in transport and communications fosters growth. This result is confirmed by Sanchez-Robles (1998) in the case of physical infrastructure and by Easterly (2001) as well as Loayza et al. (2003) in the case of communications (telephone density). On the other hand, it is argued that sometimes the inefficiency of infrastructure provision can curb and even reverse the sign of its impact on long-run growth (Devarajan et al., 1996; Hulten, 1996; Esfahani and Ramirez, 2002).

Turning to the effects on income inequality, the issue is whether infrastructure has a disproportionate impact on the income and welfare of the poor (World Bank, 2003). The presence of a disproportionately positive impact finds some support in the evidence surveyed by Brenneman and Kerf (2002). Several studies point at the effects of infrastructure on human capital accumulation: better transportation and safer roads promote school attendance; electricity allows more time for study and the use of computers; access to water and sanitation reduces child and maternal mortality. Infrastructure also connects poor people in underdeveloped areas to core economic activities, thus expanding their employment opportunities (Estache, 2003). Finally, better infrastructure in poorer regions reduces production and transaction costs (Gannon and Liu, 1997). Overall, existing studies show that infrastructure is important for economic growth and income inequality. The exact impact may depend, however, on the type of infrastructure. In the words of Sugolov et al. (2003): 'All in all, there is a broad consensus that . . . infrastructure is a necessary but not sufficient ingredient of economic growth, and that the efficient supply of the right kind of infrastructure (material and institutional) in the right place is more important than the amount of money disbursed or the pure quantitative infrastructure capacities created' (p. 3).

The aim of the present chapter is to discuss the foregoing issues from the specific point of view of 'new economic geography' (NEG), an approach to economic geography firmly grounded in recent developments in mainstream industrial organization and international trade theory.[2] NEG explains the evolution of the economic landscape as a self-organizing process driven by pecuniary externalities whose relative intensity depends on a set of well-defined microeconomic parameters. Among these, the obstacles to the geographical mobility of goods and factors are of crucial importance and can be readily related to infrastructure efficiency.

The focus of our analysis is on the effects of infrastructure on the costs of exchanging goods ('transport costs') and ideas ('communication costs'). In both cases, infrastructure can be thought of as having 'border effects' (that is, effects on the exchanges between regions) as well as 'behind-the-border effects' (that is, effects on the exchanges within regions). The result of our chapter is a unified framework that summarizes the main insights of NEG on the relation between infrastructure, economic growth and agglomeration: (1) there is a trade-off between growth and regional equality as improved infrastructure in developed 'core' regions fosters agglomeration and growth, which are instead hampered by improved infrastructure in developing 'peripheral' regions; (2) better inter-regional connections may increase rather than decrease regional inequality as improved transport and communication infrastructure between core and peripheral regions may foster not only growth but also agglomeration. The chapter is organized as follows. Section 5.2 presents the endogenous growth model that will be used as the theoretical framework. Section 5.3 discusses the feedback from growth to agglomeration. Section 5.4 highlights the effects of agglomeration on growth. Section 5.5 studies the effects of transportation and communication infrastructure on agglomeration and growth. Section 5.6 concludes.

5.2 Theoretical framework

We follow Martin and Ottaviano (1999, 2001) in modelling a spatial economy where long-run growth is sustained by ongoing product innovation and knowledge spillovers.[3] There are two regions, North and South, with the same given number Q of workers. Workers are geographically immobile and each supplies one unit of labor inelastically so that Q also measures the regional endowment of labour. Regions are also endowed with an identical initial stock of knowledge capital K_0.[4] Knowledge capital is accumulated through profit-seeking innovation performed by research and development (R&D) laboratories and is freely mobile between regions. Laboratories finance their efforts by selling bonds to workers in a perfect interregional capital market and we call $r(t)$ the riskless return on those bonds. For the sake of parsimony, in the presentation, we focus on the North with the implicit understanding that symmetric expressions apply to the South.

Consumption

Workers' preferences are defined over the consumption of two goods, a homogeneous 'traditional' good Y and a horizontally differentiated 'modern' good D. Their preferences are captured by the following utility function:

$$U = \int_{t=0}^{\infty} \log[D(t)^{\alpha} Y(t)^{1-\alpha}] e^{-\rho t} dt \tag{5.1}$$

where

$$D(t) = \left[\int_{i=0}^{N(t)} D_i(t)^{1-1/\sigma} di \right]^{1/(1-1/\sigma)}, \ \sigma > 1 \tag{5.2}$$

is the constant elasticity of substitution (CES) consumption basket of the different varieties of good D. In (5.2) $D_i(t)$ is the consumption of variety i and $N(t)$ the total number of varieties available in the economy. Instantaneous utility maximization implies that in each period workers allocate a share α of their individual expenditure $E(t)$ to the consumption of the modern good and the complementary share $1 - \alpha$ to the consumption of the traditional good. The share $\alpha E(t)$ is then distributed across varieties according to their relative prices. For any variety i, the result is individual demand:

$$D_i(t) = \frac{p_i(t)^{-\sigma}}{P(t)^{1-\sigma}} \alpha E(t) \tag{5.3}$$

where:

$$P(t) = \left[\int_{i=0}^{N(t)} p_i(t)^{1-\sigma} di \right]^{1/(1-\sigma)} \tag{5.4}$$

is the exact price index associated with the CES consumption basket (5.2) and σ is both the own- and cross-price elasticity of demand. Intertemporal utility maximization finally determines the evolution of expenditures according to a standard Euler equation:

$$\frac{\dot{E}(t)}{E(t)} = r(t) - \rho \tag{5.5}$$

where we have used the fact that (5.1) exhibits unit elasticity of intertemporal substitution.

Production

The traditional good is produced under perfect competition and constant returns to scale using labor as its only input. The unit input requirement is set to 1 by choice of units for labor, so the profit-maximizing price of Y equals the wage. Moreover, the traditional good is freely traded both between and within regions, which implies that its price and therefore its wage are the same in both regions.[5] By choosing good Y as numeraire, the common wage is also pinned down to 1.

The varieties of the modern good are produced under monopolistic competition and increasing returns to scale due to fixed and variable costs. Fixed costs are incurred in terms of knowledge capital, one unit per variety, and variable costs in terms of labor, β units per unit of output. Accordingly, in any instant t the global capital stock $K^w(t)$ determines the total number of varieties available in the economy. As in equilibrium each variety is produced by one and only one firm, $K^w(t)$ also determines the total number of firms. However, as knowledge capital is freely mobile, where varieties are actually produced is endogenously determined by the entry decisions of firms and we call $n(t)$ the number of firms producing in North. In any instant there is a large number of potential entrants that need knowledge capital to start producing. As in a given instant capital supply is fixed, their bidding for capital ends up transferring all operating profits to capital owners.

Exchanges of differentiated varieties are hampered by transport costs. These are modelled as iceberg frictions that absorb part of the quantity shipped: $\tau_N > 1$ and $\tau_R > 1$ units

have to be sent by a northern firm for one unit to be delivered to a northern and to a southern customer respectively. Symmetrically, $\tau_S > 1$ and $\tau_R > 1$ units have to be shipped by a southern firm for one unit to be delivered to a southern and to a northern customer respectively. The larger the τ's, the worse the corresponding transport infrastructure. We assume that intra-regional transportation is less costly than interregional transportation but this cost advantage is more pronounced for the North:

$$\tau_N < \tau_S < \tau_R$$

which identifies North as the developed 'core' region and South as the developing 'peripheral' one. Given our assumptions on demand and technology, all firms in any market face the same constant elasticity of demand σ and the same marginal production cost β. Hence, their profit-maximizing producer price ('mill price') is the same and equal to a constant mark-up over marginal cost:

$$p = \frac{\sigma}{\sigma - 1}\beta \tag{5.6}$$

Moreover, the consumer price ('delivered price') simply reflects different transport costs:

$$p_N = p\tau_N, p_S = p\tau_S, p_R = p\tau_R \tag{5.7}$$

Accordingly, operating profits are:

$$\pi(t) = \frac{\beta x(t)}{\sigma - 1} \tag{5.8}$$

where $x(t)$ is firm's output inclusive of the quantity absorbed by the iceberg frictions.

Finally, given (5.7), the price index (5.4) can be rewritten as:

$$P(t) = pN(t)^{\frac{1}{1-\sigma}}[\delta_N \gamma(t) + \delta_R(1 - \gamma(t))]^{\frac{1}{1-\sigma}} \tag{5.9}$$

where $\gamma(t) = n(t)/N(t)$ is the share of firms located in the North and $N(t) = K^w(t)$ is the global number of firms as well as the global stock of knowledge capital. The parameters $\delta_N = (\tau_N)^{1-\sigma}$ $(\delta_S = (\tau_S)^{1-\sigma})$ and $\delta_R = (\tau_R)^{1-\sigma}$ measure the efficiency of transportation within and between regions respectively. They are bounded between zero and one, and ranked $\delta_N > \delta_S > \delta_R$.

Innovation

The global capital stock $K^w(t)$ is accumulated through profit-seeking R&D. This is performed by perfectly competitive laboratories under constant returns to scale. In the long run, ongoing innovation is sustained by knowledge spillovers that increase the productivity of researchers as knowledge accumulates.

Martin and Ottaviano (1999, 2001) highlight two main channels through which firms' location can affect the cost of innovation: localized knowledge spillovers and intermediate business services. A general specification of the R&D technology that encompasses both is the following constant returns to scale production function:

$$\dot{K}(t) = A(t)\left[\frac{D(t)}{\varepsilon}\right]^{\varepsilon}\left[\frac{Q_I(t)}{1-\varepsilon}\right]^{1-\varepsilon} \tag{5.10}$$

where $\dot{K}(t) = dK(t)/dt$ is the flow of knowledge created at time t, $Q_I(t)$ is labor employed in R&D, and $D(t)$ is the basket of business services. This is assumed to be the same as the consumption basket for analytical convenience. Then, $0 < \varepsilon < 1$ is the share of business services in R&D production. The term $A(t)$ is total factor productivity in R&D, which is affected by knowledge spillovers. In particular, we assume that $A(t) = A\,K^{w}(t)^{\mu}[\omega_N\gamma + \omega_R(1-\gamma)]^{\mu}$ where A is a positive constant. Accordingly, $A(t)$ is an increasing function of the global stock of knowledge $K^{w}(t)$ as embodied in the activities of modern firms. The positive parameter μ measures the intensity of the knowledge spillover. The diffusion of knowledge is, however, hampered by communication costs with frictional decay regulated by the ω's. These are positive and smaller than one: ω_N measures the knowledge diffusion from northern firms to northern laboratories, and ω_R the knowledge diffusion from southern firms to northern laboratories. The larger the ω's, the better the corresponding communication infrastructure. As in the case of transportation, we assume that communication is more efficient within than between regions and this gap is more pronounced for the North:

$$\omega_R < \omega_S < \omega_N$$

The marginal cost associated with the R&D technology (5.10) is equal to:

$$F(t) = \frac{P(t)^{\varepsilon}w^{1-\varepsilon}}{A(t)} = \frac{\eta}{N(t)[\omega_N\gamma(t) + \omega_R(1-\gamma(t))]^{1-\frac{\varepsilon}{\sigma-1}}[\delta_N\gamma(t) + \delta_R(1-\gamma(t))]^{\frac{\varepsilon}{\sigma-1}}} \tag{5.11}$$

where $\eta = p^{\varepsilon}/A$ is a positive constant and we have used (5.9) as well as the fact that the equilibrium wage equals one. We have also constrained parameters so that in the long run the spatial economy develops along a balanced growth path, namely $\mu + \varepsilon/(\sigma-1) = 1$. This ensures that the marginal cost of innovation decreases through time at the same rate as its benefit as expressed by the value of (newly created) firms, thus preserving the incentive to invest in R&D (more on this in section 5.4).

As we will verify, thanks to its better local transport infrastructure, the North is the larger market. Given the rankings of ω's and δ's, (5.11) implies that the marginal cost of innovation is lower there. Therefore, given perfect competition in R&D, in equilibrium the North will attract all laboratories, so that long-run growth will be entirely driven by northern innovation.[6] Innovation in our model will be nonetheless financed in a global capital market by both northern and southern workers, which implies that in equilibrium the value $v(t)$ of a unit of knowledge capital obeys the following arbitrage condition:

$$r(t) = \frac{\dot{v}(t)}{v(t)} + \frac{\pi(t)}{v(t)} \tag{5.12}$$

This condition requires the bond yield $r(t)$ to match the percentage return on investment in knowledge capital, which consists of the percentage capital gain $\dot{v}(t)/v(t)$ and the percentage dividend $\pi(t)/v(t)$ as each unit of knowledge gives the right to the operating profits of a modern firm. Profit maximization by perfectly competitive labs finally implies that knowledge capital is priced at marginal cost: $v(t) = F(t)$.

5.3 Growth affects location

In equilibrium the arbitrage condition (5.12) implies that all firms generate the same operating profits independently from their actual locations. Given (5.8), that requires all firms to reach the same scale of output, $x(t)$, independently from their locations. Accordingly, using (5.3) and (5.7), the market clearing conditions for northern and southern firms can be written as:

$$\begin{aligned} x(t) &= \frac{p^{-\sigma}\delta_N}{P(t)^{1-\sigma}}[\alpha E(t)Q + \varepsilon F(t)\dot{N}(t)] + \frac{p^{-\sigma}\delta_R}{P^*(t)^{1-\sigma}}\alpha E^*(t)Q \\ x^*(t) &= \frac{p^{-\sigma}\delta_S}{P^*(t)^{1-\sigma}}\alpha E^*(t)Q + \frac{p^{-\sigma}\delta_R}{P(t)^{1-\sigma}}[\alpha E(t)Q + \varepsilon F(t)\dot{N}(t)] \end{aligned} \tag{5.13}$$

where variables with an asterisk pertain to the South. The asymmetry between the two conditions comes from the fact that the R&D sector is active only in the North. Since R&D demands varieties as intermediate business services, northern demand is augmented by intermediate expenditure $\varepsilon F(t)\dot{N}(t)$.

From now on, let us define the growth rate of knowledge capital as $g = \dot{K}^w(t)/K^w(t) = \dot{N}(t)/N(t)$. Moreover, to alleviate notation, let us drop the explicit dependence of variables on time when this does not generate confusion. Then, using (5.6) and (5.9), the market clearing conditions (5.13) can be solved together to yield the implied output scale:

$$x = \frac{\sigma - 1}{\beta\sigma}\frac{2\alpha EQ + \varepsilon FNg}{N} \tag{5.14}$$

and the associated location of firms:

$$\gamma = \frac{1}{2} + \frac{1}{2}\frac{\delta_R(\delta_N - \delta_S)}{(\delta_N - \delta_R)(\delta_S - \delta_R)} + \frac{\delta_N\delta_S - \delta_R^2}{(\delta_N - \delta_R)(\delta_S - \delta_R)}\left(\theta - \frac{1}{2}\right) \tag{5.15}$$

where:

$$\theta = \frac{\alpha EQ + \varepsilon FNg}{2\alpha EQ + \varepsilon FNg} \tag{5.16}$$

is the share of modern sector expenditures accruing to northern firms. Since regions share the same initial endowments, we have also set $E = E^*$. Expression (5.15) shows that the North hosts a larger number of firms because it is larger ($\varepsilon FNg > 0$ implies $\theta > 1/2$) and because it has a better intra-regional transport infrastructure ($\delta_N > \delta_S$). Both effects are amplified by any improvement in the interregional transport infrastructure (larger δ_R). Hence, we have:

Result 1: For a given growth rate, firms are attracted to the region offering larger local demand. Any improvement in the interregional transport infrastructure strengthens such attraction.

Result 2: For a given growth rate, firms are attracted to the region offering better intra-regional transport infrastructure. Any improvement in the interregional transport infrastructure strengthens such attraction.

Moreover, as θ is an increasing function of g, we can also write:

Result 3: For given expenditure, faster growth strengthens the attraction of firms to the region offering larger local demand. Any improvement in the interregional transport infrastructure magnifies such effect.

This shows that growth affects location. In particular, agglomeration is an increasing function of the growth rate.

5.4 Location affects growth

To characterize the long-run growth of the economy, we focus on a balanced path along which expenditure as well as the growth rate are constant. With constant expenditure, $\dot{E} = 0$ so that (5.5) gives $r = \rho$. Since, by (5.11) and (5.15), also FN and γ are constant, the evolution of the value of knowledge capital is driven by the growth rate through the implied change in the marginal cost of R&D, $\dot{v}/v = \dot{F}/F = -g$, which shows that the marginal cost (F) and the marginal benefits of innovation (v) both fall at the same constant rate. Then, by (5.8) and (5.14), the arbitrage condition (5.12) can be rewritten as:

$$\rho = -g + \frac{2\alpha EQ + \varepsilon FNg}{\sigma FN} = \frac{2\alpha EQ}{\sigma FN} - g\left(\frac{\sigma - \varepsilon}{\sigma}\right) \tag{5.17}$$

The model is closed by imposing the labour market clearing condition whereby the global endowment of labor $2Q$ is fully employed in innovation $Q_I = (1-\varepsilon)FNg$, in modern production $Q_D = [(\sigma - 1)/\sigma][2\alpha EQ + \varepsilon FNg]$, and in traditional production $Q_Y = 2(1-\alpha)EQ$. Simplification leads to the full employment condition:

$$2Q = \frac{\sigma - \varepsilon}{\sigma} FNg + 2\frac{\sigma - \alpha}{\sigma} EQ \tag{5.18}$$

Solving (5.17) together with (5.18) shows that in equilibrium expenditure equals permanent income:

$$2EQ = 2Q + \rho FN \tag{5.19}$$

where $2Q$ is labour income and ρFN is the additional income from ('annuity value' of) the initial stock of knowledge capital. Accordingly, both terms on the right-hand side of (5.19) are evenly split between regions.[7]

By (5.17) or (5.18) the corresponding growth rate is:

$$g = \frac{\alpha}{\sigma - \varepsilon}\frac{2Q}{FN} - \rho\frac{\sigma - \alpha}{\sigma - \varepsilon} \tag{5.20}$$

which shows that location affects growth through the cost of innovation FN.[8] In particular, given (5.11), more agglomeration in the North makes innovation cheaper and leads to faster growth. The more so, the better northern infrastructure is with respect to the interregional one. Accordingly, we can write that:

Result 4: Further agglomeration in the core region offering better intra-regional transport and communication infrastructure fosters growth.

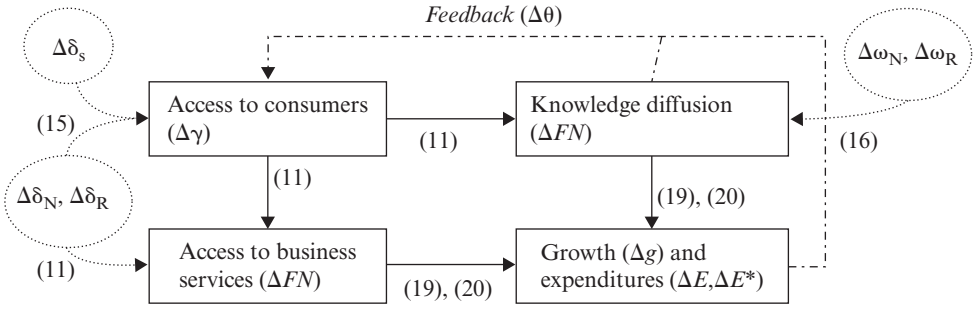


Figure 5.1 Infrastructure, location and growth

Moreover, in (5.20) the growth rate is a function of ε, which measures how much business services matter for R&D with respect to labour. A larger ε has three effects on growth. Two effects are direct and promote growth. The third is indirect, works through the cost of innovation and is ambiguous in sign. First, if R&D becomes relatively more intensive in business services than in labour, a higher demand for differentiated inputs by the innovation sector increases firms' operating profits. Along a balanced growth path, g accelerates to keep capital markets in equilibrium (see 5.17). Second, from the labour market equilibrium, as ε goes up, the number of workers hired in the modern sector is not enough to compensate for those fired in the innovation sector, unless g goes up too (see 5.18). Third and last, if business services become more important for R&D, the cost of innovation may rise or not depending on the relative efficiency of communication with respect to transportation (see 5.11). For this reason, the overall impact of larger ε on growth is ambiguous.

5.5 Infrastructure, agglomeration and growth

The equilibrium of the model is fully characterized by expressions (5.11), (5.15), (5.16), (5.19) and (5.20). It therefore features complex interactions between changes in location ($\Delta\gamma$) and growth (Δg) filtered through changes in expenditures (ΔE) and R&D costs (ΔFN). These interactions are summarized graphically in Figure 5.1.

Abstracting from the feedback (5.16) going from growth, R&D costs and expenditure to location, changes in intra-North and interregional communication infrastructure ($\Delta\omega_N$, $\Delta\omega_R$) affect the diffusion of knowledge (ΔFN). In particular, better intra-North and interregional communication reduces the cost of innovation ($\Delta FN < 0$) and fosters growth ($\Delta g > 0$). Changes in intra-South communication infrastructure ($\Delta\omega_S$) are instead irrelevant as no R&D takes place in the South. This does not hold for changes in intra-South transport infrastructure ($\Delta\delta_S$). In particular, any improvement ($\Delta\delta_S > 0$) attracts firms to the South ($\Delta\gamma < 0$), which increases the cost of innovation ($\Delta FN > 0$) and slows growth ($\Delta g < 0$). On the other hand, better intra-North and interregional transportation attracts firms to the North ($\Delta\gamma > 0$), which decreases the cost of innovation ($\Delta FN < 0$) and fosters growth ($\Delta g > 0$). The overall picture can be gauged by substituting (5.11) into (5.20), which allows us to write growth as function of location:

$$g = \frac{\alpha}{\sigma - \varepsilon}\,\frac{2Q}{\eta}[\omega_N\gamma + \omega_R(1-\gamma)]^{1-\frac{\varepsilon}{\sigma-1}}[\delta_N\gamma + \delta_R(1-\gamma)]^{\frac{\varepsilon}{\sigma-1}} - \rho\frac{\sigma-\alpha}{\sigma-\varepsilon} \qquad (5.21)$$

Turning to the feedback (5.16) implies that changes in the cost of innovation (ΔFN), growth (Δg) and expenditure (ΔE) affect the share of northern demand ($\Delta\theta$) and thus firms' location ($\Delta\gamma$). Specifically, using (5.19) and (5.20), the share of northern demand (5.16) can be rewritten as:

$$\theta = \frac{1}{2} + \frac{1}{2}\frac{\varepsilon}{\sigma}\frac{g}{g+\rho} \tag{5.22}$$

and hence (5.15) as:

$$\gamma = \frac{1}{2} + \frac{1}{2}\frac{\delta_R(\delta_N - \delta_S)}{(\delta_N - \delta_R)(\delta_S - \delta_R)} + \frac{1}{2}\frac{\delta_N\delta_S - \delta_R^2}{(\delta_N - \delta_R)(\delta_S - \delta_R)}\frac{\varepsilon}{\sigma}\frac{g}{g+\rho} \tag{5.23}$$

which shows location as a function of growth through its influence on the share of expenditure.

The foregoing expressions highlight the crucial result of our NEG framework: there is 'circular causation' between agglomeration (larger γ) and growth (larger g). As growth fosters agglomeration and agglomeration fosters growth, policy-makers face a trade-off between promoting growth and challenging regional inequality. Given Results 1 and 2, the implication is:

> **Result 5**: Policies designed to improve interregional and intra-core infrastructure foster agglomeration in the core region as well as growth. Policies designed to improve intra-periphery infrastructure foster relocation from the core to the peripheral regions but hamper growth.

Expression (5.23) also shows that the feedback from growth to agglomeration gets stronger the more R&D relies on business services (the larger ε). By (5.21), the impact of ε on the effect of agglomeration on growth is, instead, ambiguous.

Finally, taken together, (5.21) and (5.23) implicitly define the equilibrium values of g and γ along the balanced growth path. As both are highly non-linear, they are not amenable to explicit analytical solution. One way to gain extra analytical insight is to focus on the two scenarios analysed by Martin and Ottaviano (1999) and Martin and Ottaviano (2001). These can be retrieved from our framework as specific polar cases that arise when the cost of innovation is respectively affected by communication costs only ($\varepsilon = 0$) and by transport costs only ($\varepsilon = \sigma - 1$).

Communication costs

When $\varepsilon = 0$, the cost of innovation does not depend on transport costs as business services are not used in R&D. Accordingly, (5.11) becomes simply:

$$FN = \frac{\eta}{\omega_N\gamma + \omega_R(1-\gamma)} \tag{5.24}$$

In this case, since by (5.23) the expenditure share is the same in both regions ($\theta = 1/2$), location is unaffected by growth:

$$\gamma = \frac{1}{2} + \frac{1}{2}\frac{\delta_R(\delta_N - \delta_S)}{(\delta_N - \delta_R)(\delta_S - \delta_R)} \tag{5.25}$$

where $\delta_N/(\delta_N - \delta_R) > \delta_R/(\delta_S - \delta_R)$ has to be imposed to concentrate on the meaningful case in which at least some firms locate in the South.

On the contrary, growth is affected by location as (5.20), (5.24) and (5.25) together imply:

$$g = \frac{\alpha}{\sigma}\frac{Q}{\eta}\left[(\omega_N + \omega_R) + (\omega_N - \omega_R)\frac{\delta_R(\delta_N - \delta_S)}{(\delta_N - \delta_R)(\delta_S - \delta_R)}\right] - \rho\frac{\sigma - \alpha}{\sigma} \tag{5.26}$$

Hence, by inspection of (5.26), we get:

Result 6: When the cost of innovation is affected by communication costs only, improved communication infrastructure within the core region fosters growth and has no impact on agglomeration. The same applies to improved interregional communication. Differently, improved communication within the peripheral region has no impact whatsoever as long as no innovation takes place there. Changes in transport infrastructure affect location but have no impact on growth.

Transportation costs

When $\varepsilon = \sigma - 1$, the cost of innovation is affected by transport costs but not by communication costs. In this case, knowledge spillovers are 'global' since all laboratories benefit from the global stock of knowledge capital in the same way, independently from their actual locations. Expressions (5.11) and (5.22) respectively become:

$$FN = \frac{\eta}{\delta_N\gamma + \delta_R(1 - \gamma)} \tag{5.27}$$

$$\gamma = \frac{1}{2} + \frac{1}{2}\frac{\delta_R(\delta_N - \delta_S)}{(\delta_N - \delta_R)(\delta_S - \delta_R)} + \frac{1}{2}\frac{\delta_N\delta_S - \delta_R^2}{(\delta_N - \delta_R)(\delta_S - \delta_R)}\frac{\sigma - 1}{\sigma}\frac{g}{g + \rho} \tag{5.28}$$

which shows that, even with global spillovers, when R&D uses business services, location is affected by growth. On the other hand, setting $\varepsilon = \sigma - 1$ in (5.21) also shows that location affects growth:

$$g = \alpha\frac{2Q}{\eta}\left[\delta_N\gamma + \delta_R(1 - \gamma)\right] - \rho(\sigma - \alpha) \tag{5.29}$$

The two conditions (5.28) and (5.29) can be respectively visualized as a concave and a linear increasing functions mapping from g to γ. To see this, simply invert (5.29) to express γ as a function of g:

$$\gamma = \frac{g + \rho(\sigma - \alpha)}{2\frac{\alpha}{\eta}Q(\delta_N - \delta_R)} - \frac{\delta_R}{\delta_N - \delta_R} \tag{5.30}$$

An increase in δ_S has no impact on (5.30) as no R&D takes place in the South. Instead, it shifts (5.28) downwards, which reduces agglomeration and growth. Moreover, most naturally, what we proved in the general case also holds in this specific case: increases in δ_N and δ_R foster both agglomeration and growth. Hence, we can write:

Result 7: When the cost of innovation is affected by transport costs only, improved transportation infrastructure within the core region and between regions fosters agglomeration and growth whereas improved transportation infrastructure in the

periphery hampers them. Changes in communication infrastructure have no impact whatsoever.

5.6 Conclusion

We have proposed a simple theoretical framework to study the impact of infrastructure on economic growth and regional imbalances. The framework has presented in a unified way the main insights of NEG models with endogenous growth and free capital mobility. Two main results stand out. First, there is a trade-off between growth and regional equality as improved infrastructure in developed 'core' regions fosters both agglomeration and growth, which are instead hampered by improved infrastructure in developing 'peripheral' regions. Second, better interregional connections may increase rather than decrease regional inequality as improved transport and communication infrastructure between core and peripheral regions fosters not only growth but also agglomeration. These insights are confirmed by Fujita and Thisse (2003) when also labour is mobile. These authors stress the fact that increased agglomeration does not necessarily imply the impoverishment of peripheral regions as long as its positive impact on growth is strong enough.

Two caveats are in order. On the one hand, our results hold no matter whether improvements involve transportation or communication. This equivalence does not survive the introduction of barriers to capital mobility. In that case, Baldwin et al. (2001) show that improved transportation increases regional inequality whereas improved communication decreases it. On the other hand, as suggested by Fujita and Mori (2005), endogenous growth theories embodied in NEG have so far assumed some kind of ad hoc knowledge spillovers or externalities without providing their micro-foundations. Our framework is no exception and that issue surely deserves further attention by future research.

Notes

1. Financial support from MIUR and the University of Bologna is gratefully acknowledged. We have benefited from comments by an anonymous referee.
2. There exist many surveys of NEG and alternative approaches to spatial issues. See, for example, Fujita et al. (1999), Fujita and Thisse (2002), Baldwin et al. (2003).
3. See Baldwin and Martin (2004) for a broader survey of NEG models with endogenous growth.
4. Assuming identical factor endowments across regions allows us to highlight the specific role of infrastructure in determining regional specialization.
5. The assumption that the traditional good is freely traded is rather standard in NEG models. However, it is not innocuous. For instance, the computable general equilibrium analysis by Kilkenny (1998) suggests that the location of the modern sector reacts differently to improved transportation depending on the relative importance of the transport costs on the modern and the traditional goods. See also Chapter 7 in Fujita et al. (1999) on the same topic.
6. In principle, it is possible to imagine alternative configurations for transportation and communication costs. For example, Hirose and Yamamoto (2007) set $\varepsilon = 0$, and $\delta_S = \delta_N = \omega_S = \omega_N = 1$. They also allow for asymmetric interregional communication costs. For instance, firms in the South may benefit more than firms in the North from interregional knowledge spillovers. If this advantage in absorptive capacity is strong enough, the South (the smaller market) will end up hosting the innovation sector, and this will affect the relations among agglomeration, growth and regional inequality. However, it can be argued that, even under asymmetric interregional communication costs that favour knowledge absorption in the South, allowing $\varepsilon > 0$, as we do in our chapter, makes the location of innovation in the North more likely.
7. Since regional incomes are equalized in nominal terms, disparities in real incomes are driven by firm location and different intra-regional transport costs.
8. *FN* is the marginal cost of innovation net of the spillover from accumulated knowledge capital.

References

Aschauer, D. (1989), 'Is public expenditure productive?', *Journal of Monetary Economics*, **23**, 177–200.

Baldwin, R., and P. Martin (2004), 'Agglomeration and regional growth', in J.V. Henderson and J.-F. Thisse (eds), *Handbook of Urban and Regional Economics*, Vol. 4, New York: North-Holland, pp. 2671–711.

Baldwin, R., P. Martin and G.I.P. Ottaviano (2001), 'Global income divergence, trade and industrialization: the geography of growth take-off', *Journal of Economic Growth*, **6**, 5–37.

Baldwin, R., R. Forslid, P. Martin, G.I.P. Ottaviano and F. Robert-Nicoud (2003), *Economic Geography and Public Policy*, Princeton, NJ: Princeton University Press.

Brenneman, A. and M. Kerf (2002), 'Infrastructure and poverty linkages: a literature review', World Bank, mimeo.

Calderon, C.A. and L. Serven (2003), 'The output cost of Latin America's infrastructure gap', in W. Easterly and L. Serven (eds), *The Limits of Stabilization: Infrastructure, Public Deficits, and Growth in Latin America*, Stanford, CA: Stanford University Press and the World Bank, pp. 95–118.

Calderon, C.A. and L. Serven (2004), 'The effects of infrastructure development on growth and income distribution', World Bank Policy Research Working Paper No. 3400.

Devarajan, S., V. Swaroop and H. Zhou (1996), 'The composition of public expenditure and economic growth', *Journal of Monetary Economics*, **37**, 313–44.

Easterly, W. (2001), 'The lost decade: developing countries' stagnation in spite of policy reform', World Bank, mimeo.

Easterly, W. and S. Rebelo (1993), 'Fiscal policy and economic growth: an empirical investigation', *Journal of Monetary Economics*, **32**, 417–58.

Esfahani, H. and M.T. Ramirez (2002), 'Institutions, infrastructure and economic growth', *Journal of Development Economics*, **70**, 443–77.

Estache, A. (2003), 'On Latin America's infrastructure privatization and its distributional effects', World Bank, mimeo.

Fujita, M. and T. Mori (2005), 'Frontiers of the new economic geography', *Papers in Regional Science*, **84**, 377–405.

Fujita, M. and J.-F. Thisse (2002), *Economics of Agglomeration*, Cambridge: Cambridge University Press.

Fujita, M. and J.-F. Thisse (2003), 'Does geographical agglomeration foster economic growth? And who gains and loses from it?', *Japanese Economic Review*, **54**, 121–45.

Fujita, M.,P. Krugman and A. Venables (1999), *The Spatial Economy*, Cambridge, MA: MIT Press.

Gannon, C. and Z. Liu (1997), 'Poverty and transport', World Bank, mimeo.

Gramlich, E. (1994), 'Infrastructure investment: a review essay', *Journal of Economic Literature*, **32**, 1176–96.

Hirose, K. and K. Yamamoto (2007), 'Knowledge spillovers, location of industry, and endogenous growth', *Annals of Regional Science*, **41**, 17–30.

Hulten, C. (1996), 'Infrastructure capital and economic growth: how well you use it may be more important than how much you have', NBER Working Paper 5847.

Kilkenny, M. (1998), 'Transport costs and rural development', *Journal of Regional Science*, **38**, 293–312.

Loayza, N., P. Fajnzylber and C. Calderon (2003), 'Economic growth in Latin America and the Caribbean: stylized facts, explanations and forecasts', World Bank, mimeo.

Martin, P. and G.I.P. Ottaviano (1999), 'Growing locations: industry location in a model of endogenous growth', *European Economic Review*, **43**, 281–302.

Martin, P. and G.I.P. Ottaviano (2001), 'Growth and agglomeration', *International Economic Review*, **42**, 947–68.

Röller, L-H. and L. Waverman (2001), 'Telecommunications infrastructure and economic development: a simultaneous approach', *American Economic Review*, **91**, 909–23.

Sanchez-Robles, B. (1998), 'Infrastructure investment and growth: some empirical evidence', *Contemporary Economic Policy*, **16**, 98–108.

Sugolov, P., B. Dodonov and C. von Hirschhausen (2003), 'Infrastructure policies and economic development in East European transition countries: first evidence', DIW Berlin WP-PSM-02.

World Bank (1994), *World Development Report 1994: Infrastructure for Development*, Washington, DC: World Bank.

World Bank (2003), *Inequality in Latin America and the Caribbean*, Washington, DC: World Bank Latin American and Caribbean Studies.

PART II

DEVELOPMENT THEORIES: REGIONAL PRODUCTION FACTORS

6 Agglomeration, productivity and regional growth: production theory approaches

Jeffrey P. Cohen and Catherine J. Morrison Paul

6.1 Introduction

One definition of agglomeration economies is that 'cost reductions occur *because* economic activities are *located* in one place' (McDonald and McMillen, 2007), an idea typically attributed to Marshall (1920). Ohlin (1933) more specifically categorized agglomeration economies by distinguishing localization economies and urbanization economies, which has become a standard in the urban economics literature. Localization economies involve benefits to firms from expansion of their own industry, resulting in industry 'clusters'. Urbanization economies occur when expansion of an urban area benefits firms from the proximity of a variety of industries, leading to regional growth.

Both of these types of agglomeration economies, arising respectively from geographic concentration of 'specialized' and 'diverse' production, imply lower costs in real terms rather than in nominal terms for firms. However, because they result from factors beyond the control of individual firms, agglomeration economies are often theoretically modeled as external scale economies.

The 'causes' of agglomeration economies may take a variety of forms. For example, proximity of 'like' firms may increase the quantity or quality of the labor pool so better matches or less risk are involved in hiring; proximity of suppliers may involve easier access to or lower costs of materials inputs; or proximity of population concentrations may facilitate distribution of products. These specific channels explaining clustering, as has been discussed at length in the agglomeration economies literature (and summarized by Rosenthal and Strange, 2004), are often called the 'micro-foundations' underlying agglomeration economies.

The 'effects' of these external factors theoretically involve shifts of a firm's production or cost curves. These effects may generally be thought of as arising from availability or augmentation of some sort of input, or a more general 'disembodied' enhancement of production possibilities from the proximity of other productive activity. That is, factors external to a particular firm but associated with firm density or clustering increase firm productivity, implying more output for a given amount of inputs or less input cost to produce a given amount of output. The enhanced economic performance of firms from agglomeration economies in turn results in regional growth from firms' location choices.

In this chapter we discuss the empirical representation of agglomeration economies, with a focus on the potential of production theory-based econometric models to analyze the productive impacts of such externalities. In particular, we overview the use of production theory models and measures to represent the causes, and productivity, location and growth effects, of agglomeration economies.

6.2 The conceptualization of agglomeration economies: causes and effects

The two general types of agglomeration economies typically distinguished are economies from own-industry concentration, or localization economies, and from density of all economic activity in a particular area, or urbanization economies. Issues arise about the industrial variety (specialization versus diversity), geographic distance and temporal dynamics involved in agglomeration economies. However, the fundamental point is that proximity of productive activity in own or other industries confers external benefits on firms that enhance their economic performance and thus motivate clustering.[1] Empirical analysis typically involves measuring the economic performance or growth impacts of such externalities.

Many possible causes of agglomeration economies have been identified in the literature. The characterization of these causes is typically into three 'Marshallian' channels through which agglomeration works – labor market pooling, input sharing and knowledge spillovers – all of which involve (external and possibly also internal) economies of scale. Duranton and Puga (2001) argue that a combination of these agglomeration drivers, each of which can be considered a form of 'sharing', is likely to be prevalent in any geographically concentrated area.

Labor market pooling occurs when workers can easily move between clustered firms in an industry, which lowers job search costs and risk, and facilitates hiring. As noted by O'Sullivan (2007), this implies a more elastic labor supply curve for clustered relative to isolated firms, so they are able to respond to low versus high labor demand periods by adjusting employment with little pressure on wages. A larger pool of workers also facilitates better skill 'matches' between workers and employers, possibly resulting in specialization (Baumgardner, 1988). Overall, labor market pooling confers positive externalities due to ready access of both employers and employees to alternatives.

Input sharing implies that density of productive activity allows firms to outsource their inputs from suppliers who can produce at an efficient scale of production by exploiting (internal) scale economies.[2] Firms with greater purchased input intensity will thus benefit more from locating close to input suppliers (Holmes, 1999).[3] This is similar to the notion of supplier (versus customer) concentration or 'thickness' conferring cost savings on firms (Bartlesman et al., 1994; Morrison and Siegel, 1999).

Knowledge spillovers mean that interactions among entrepreneurs or workers, which are facilitated by high firm density, enhance economic performance. Such spillovers are difficult to capture empirically because they typically do not involve purchased inputs. However they are often assumed to be related to labor skill (human capital) or research and development (R&D) intensity (Audretsch and Feldman, 1996; Morrison and Siegel, 1998). Although some have suggested that in modern times geographic proximity should not affect knowledge transmission (Krugman, 1991), others have emphasized that knowledge, unlike information, is best conveyed through physical proximity (Audretsch, 1998; Von Hipple, 1994; Glaeser et al., 1992).

Other proposed causes of firms' geographic density include concentration of demand and natural advantage.[4] Concentration of demand implies population density, so it may be considered one aspect of urbanization, although urban effects are more typically defined in terms of economic activity in general. Demand concentration is often conceptualized in terms of home market effects (Krugman, 1980; Davis and Weinstein, 1999), where firms' density or size results in concentration of employment and thus demand that attracts other firms.[5]

Natural advantage involves factor endowments that may be generally characterized as local supplies of primary inputs, and that confer economies from, say, transportation cost savings (see Fuchs, 1962, for a general overview). For example, climate (Marshall, 1920), natural resources (Kim, 1995) or proximity of primary materials like agricultural products (Audretsch and Feldman, 1996) may encourage firms to locate in close proximity to these natural assets. An input such as labor may also fall into this category if it is immobile (Kim, 1999; Ellison and Glaeser, 1999), although this is not likely to be relevant in most modern industries (except perhaps in the short term).

This type of externality involves external inputs rather than externalities or spillovers among firms (Moreno et al., 2004). That is, natural advantage involves an input associated with a particular location, such as local natural resources that become primary materials inputs. Other factors that may act as external inputs include public capital such as transportation or communications infrastructure. The distinction between externalities and external inputs may not be empirically transparent, however, because whether agglomeration involves a spillover or a locally available factor that acts as or augments an input may depend on the mechanism through which it works. Human capital, for example, may be a local external input because the general population has more education or skills, but could also act as a conduit for knowledge spillovers.

Although these are the general categories of agglomeration causes or drivers usually identified, it is difficult to pinpoint specific causes for specific circumstances. Essentially, any factors associated with proximity or density of production and population that comprise or enhance inputs available to firms and affect productivity and growth are potential channels for agglomeration economies. Empirical representation of such factors is even more imprecise. Direct evidence of, say, knowledge spillovers may at times be available, such as the prevalence of patent citations in the same Metropolitan Statistical Area (MSA) (Jaffe et al., 1993). However, measures are usually more indirect, such as education levels proxying knowledge spillovers (Audretsch and Feldman, 1996).

In turn, the effects of such agglomeration drivers involve various aspects of economic performance from clustering such as enhanced innovation, higher input (labor or capital) demand or price (wage or asset value), greater productivity, reduced costs, and location decisions (firm 'births').[6] For example, as further discussed below, significant product introductions (as a measure of innovation) and their link to agglomeration factors were considered by Audretsch and Feldman (1996). Higher wages and their connection to education levels or human capital were targeted by Rauch (1993), Moretti (2004) and Rosenthal and Strange (forthcoming). The productivity effects of industry characteristics such as types of local employment (reflecting labor pooling) were analyzed by Henderson (2003). Cost savings and location motivations from the proximity of primary agricultural inputs were explored by Cohen and Paul (2005).

Such 'micro' analyses, which at least theoretically rely on the notion of enhanced firm productivity,[7] provide the basis for evaluation of firm location decisions and ultimately the more 'macro' notion of economic growth. Although this link is difficult to make empirically, work has proceeded on the dependence of firm 'births' on agglomeration factors (Carlton, 1983; Rosenthal and Strange, 2003; Barrios et al., 2005).[8] Such models also provide the foundation for (endogenous) growth models; greater productivity for firms in a particular location will result in regional growth.

6.3 Modeling agglomeration economies

Empirical analysis of agglomeration economies typically involves characterizing one or more of their causes by proxies and relating these to observed concentration of firms.[9] The effects of this concentration, in terms of firms' economic performance or location choices, are then measured to assess the extent of external economies or the productivity effects of agglomeration.

Questions addressed in the literature often involve absolute or relative productive impacts of the various causes of agglomeration economies (such as whether localization or urbanization economies have greater impacts; Nakamura, 1985; Henderson, 1986), or whether spillovers from one or more Marshallian causes of agglomeration economies are evident (Henderson, 2003). Econometric analysis of such questions requires specifying a theoretical model that identifies the performance or location variable of interest, its underlying arguments or determinants, and the measures that can be constructed to represent the agglomeration economies.

One branch of this literature focuses on labor demand shifts,[10] usually specified in terms of employment or wages. For example, the relationship between employment growth and concentration was targeted by Henderson et al. (1995) and Glaeser et al. (1992), assuming that enhanced productivity from agglomeration economies implies greater labor demand. Analyses of wage rate differentials such as Glaeser and Mare (2001) and Wheaton and Lewis (2002) similarly are based on the notion that enhanced productivity implies a greater marginal product of labor and thus demand and wages.

However, such analyses are partial, as noted by Rosenthal and Strange (2004), because: 'existing employers are constrained by prior choices, most importantly the level and kind of capital previously installed'. That is, existing (quasi-fixed) capital levels affect how the firm values labor on the margin, and thus how it changes its employment choices in response to external changes. In the longer run substitution possibilities with capital (as well as other inputs) will affect the amount overall productivity and employment growth or wage levels are related.

Another less prevalent but similar (single-input) approach is to target capital instead of labor demand, in the context of asset value or rents, with the idea that firms will be willing to pay higher rents only if they gain higher productivity (Rosen, 1979; Roback, 1982). Due to data limitations, implementing this idea sometimes requires using unsatisfactory proxies such as rent for housing rather than firms (Dekle and Eaton, 1999), although a detailed production theory model can facilitate representation of capital shadow values and their dependence on density (Cohen and Paul, 2007).

More generally, establishing the existence and the extent of agglomeration economies involves modeling and measuring productivity more directly. However, many studies on the productivity effects of agglomeration also focus on only one input – typically labor productivity. For example, Ciccone and Hall (1996) and Ciccone (2002) analyze the relationship between regional employment density and labor productivity growth. Although they control for education and capital intensity, this remains a partial perspective on the effects of agglomeration because productivity differentials affect demands for both labor and capital (as well as other inputs), with the balance depending on prices, substitution possibilities and other market and technological factors.

That is, single-input models are partial representations of a production theory characterization of total- or multi-factor productivity, which involves changes in marginal

products of all inputs and may be non-neutral. More specifically, in a standard production function framework productivity growth is represented as increased output from a given amount of (internal) inputs. In the context of agglomeration economies this means a production function shift from an externality or spillover associated with density of other firms (or population), which may not be parallel (neutral), but could differ for different inputs. For example, labor pooling or knowledge spillovers could have a greater impact on the 'productivity' of labor, and input sharing on capital.

Such a shift may be theoretically modeled, for example as in Rosenthal and Strange (2004), by augmenting a standard production function model as:

$$y_j = g(A_j)f(\mathbf{x}_j) \tag{6.1}$$

for firm j, where y_j denotes aggregate output, the vector $\mathbf{x}_j$ includes levels of inputs traditionally specified in production theory models (say, labor, capital, materials), and $g(A_j)$ represents production function shifts from environmental factors underlying agglomeration (external scale) economies. This is similar to models of shifts in $f(\mathbf{x}_j)$ over time from technical change, typically written in terms of a multiplicative factor $A(t)$ in microeconomics textbooks. However, the impacts of distance are somewhat more complicated than those of time, because space is not as readily defined (at least linearly).

In particular, Rosenthal and Strange 'ideally' write A_j as $A_j = q(\mathbf{x}_j,\mathbf{x}_k)a(d^G_{jk},d^I_{jk},d^T_{jk})$, where k denotes other firms for which spillovers with firm j occur. Thus, $q(\mathbf{x}_j,\mathbf{x}_k)$ reflects externalities that depend on the input levels (and scale) of firms j and k, and $a(d^G_{jk},d^I_{jk},d^T_{jk})$ captures the different dimensions along which 'distance' can be measured – spatial (G, geographic proximity, such as the same county or state), industrial (I, type of economic activity that confers externalities, such as own industry or suppliers), and temporal (T, the time dimension, such as learning with a lag). A_j can also include factors such as local availability of primary materials or infrastructure that act as external inputs.

A_j is typically, however, specified in terms of one or a limited number (somehow aggregated) of less detailed proxies for agglomeration drivers such as a general measure of density or scale. For example, in studies on the productivity effects of city expansion, as documented by Rosenthal and Strange (2004), A_j is usually specified as city size (population or employment), although it could also be defined in terms of other external spillovers or inputs (as elaborated in the next section).[11] For example, the intensity of R&D in a firm's locality, as well as its internal R&D, may generate knowledge that benefits the firm (a knowledge production function; Griliches, 1979). The former then comprises a component of an $\mathbf{A}_j$ vector whereas the latter is part of the $\mathbf{x}_j$ vector. Similarly, as noted by Audretsch (1998); 'key factors generating new economic knowledge include a high degree of human capital, a skilled labour-force, and a high presence of scientists and engineers'. This suggests including measures of local education levels and university research as $\mathbf{A}_j$ components (Baptista, 1997).

The multiplicative nature of A_j in (6.1) imposes neutrality of the productivity effect, or separability of the 'inputs' in $\mathbf{x}_j$ and A_j, which is assumed in most of the literature and supported by Henderson (1986). However, if factors in A_j have differential (non-neutral) input effects (Cohen and Paul, 2004), A_j or perhaps a vector of agglomeration causes or factors $\mathbf{A}_j$ should be included directly as arguments of the production function:

$$y_j = f(\mathbf{x}_j, \mathbf{A}_j). \tag{6.2}$$

If $f(\cdot)$ takes a flexible functional form such as a translog (second-order approximation in logarithms) or generalized Leontief (second-order approximation in square roots), this function captures the dependence of the $\mathbf{x}_j$ marginal products on both other input levels and the A_j variables (interaction or cross-effects).

Individual input (labor or capital) demand models are theoretically related to a production function such as (6.2) because increasing overall productivity from $\mathbf{A}_j$ factors implies greater marginal productivity or values of, and thus demands for, the inputs in $\mathbf{x}_j$. However, a full production function model, particularly when specified without neutrality (separability) assumptions imposed and approximated to the second order, recognizes both substitutability among inputs and input-specific shift effects. Further, in such a model the link between agglomeration economies and productivity is more direct because $\mathbf{A}_j$ factors that directly increase marginal products of inputs indirectly translate into higher wages (or asset values) and employment (or investment).

A dual production theory model such as a cost function permits an even more explicit representation of agglomeration economies, since they are conceptualized in terms of production costs – as well as a direct link to input demands. That is, assuming cost minimization, Shephard's lemma can be used to specify labor (and capital or other input) demand explicitly as a function of all cost function arguments. This will include the factors in the $\mathbf{A}_j$ vector by duality theory; input cost minimization given the production function results in a cost function that depends not only on the observed output level and prevailing input prices but also on other production function arguments.

More formally, the production function $y_j = f(\mathbf{x}_j, \mathbf{A}_j)$ is dual to the cost function

$$C_j = g(\mathbf{p}_j, y_j, \mathbf{A}_j), \tag{6.3}$$

where $\mathbf{p}_j$ is the vector of prices p_{nj} of the N inputs x_{nj} and C_j is total input costs, $\Sigma_n x_{nj} p_{nj}$. Although estimation of such a cost function permits a detailed representation of agglomeration economies and their input use and composition implications, it also imposes more data requirements than wage or even production function specifications. In particular, input prices as well as levels for inputs such as capital may be difficult to obtain. However, if data permits, production theory models such as (6.2) and (6.3) provide great potential for analysis of the causes and effects of agglomeration economies.

For example, plant-level output and input data were effectively used by Henderson (2003) to analyze agglomeration economies in a production function model, and by Chapple et al. (2006) to assess location-specific costs of waste reduction in a cost function model. Data at higher levels of aggregation may also be relevant to address many spatial issues, such as at the state or even national level (Cohen and Paul, 2004; Bartlesman et al., 1994). Recognizing and exploiting the potential of production theory models is thus an important step in the empirical literature on agglomeration economies.

6.4 Measuring agglomeration economies

The A_j variables

In addition to data on production output(s) y_j[12] and inputs $\mathbf{x}_j$, empirical analysis of agglomeration economies requires measures of their causes that comprise components of

the $\mathbf{A_j}$ vector. Although these A_j factors are typically motivated in terms of external spillovers such as knowledge sharing or external inputs such as transportation infrastructure, direct measures or even proxies of such factors are often not even conceptually (much less practically) available. For example, theoretical discussions of A_j variables that typically involve the scale and proximity of surrounding productive activity on a variety of dimensions – spatial, industry and temporal – are at least somewhat amorphous. More specific external input variables such as climate or availability of public infrastructure are perhaps more readily defined, but often are not easily measurable at an appropriate (for example, industrial or spatial) level of aggregation.

Various proxies have thus been used for such factors. Urbanization economies are often measured via city size, based on total employment or population (Nakamura, 1985; Henderson, 1986). Localization economies are similarly proxied by employment in the own industry. Specialization may be represented by employment share in a particular industry (Glaeser et al., 1992; Henderson et al., 1995), although this raises questions about levels (size) versus specialization (intensity) or absolute versus relative changes.[13] Diversity can be captured by the share of production attributed to secondary industries in a particular location (Glaeser et al., 1992), or an index of employment diversity (Henderson et al., 1995).[14]

For the Marshallian causes of agglomeration effects, labor market pooling implies better matching of employers and employees, although this is difficult to measure. One possibility is to focus on specialization, which may be greater if either matches are easier or the market is larger (Baumgardner, 1988). Lower wages could also be an indicator of close matches via a lower risk premium (Diamond and Simon, 1990). Such proxies for labor pooling, however, are somewhat limited because they could be associated with various characteristics of the location and the industry base.

Input sharing may be represented by purchased input intensity (relative to sales), which has been found to be positively related to concentration (Holmes, 1999). Holmes also suggests that input sharing is directly related to the proportion of specialized input suppliers, although this is difficult to represent empirically; he uses data on specialized finishing plants for the textiles industry. Holmes and Stevens (2002) show that establishment size may represent input sharing because it implies scale economies in the production of intermediate inputs. Bartlesman et al. (1994) and Morrison and Siegel (1999) similarly focus on concentrations of input suppliers using more aggregated data at higher and lower SIC code levels. (SIC code represents Standard Industrial Classifications, which break down production into different categories.)

Knowledge spillovers may be particularly difficult to capture empirically. Direct measures are sometimes available, such as patent citations in the same MSA (Jaffe et al., 1993). More indirect measures of knowledge 'orientation' such as university research in a particular field, ratios of R&D expenditure to sales or intensity of skilled labor are also used (Audretsch and Feldman, 1996). Education levels, as an indicator of human capital, have been related to higher wages (Rauch, 1993). Similarly, Moretti (2004) links the number of college graduates to wage levels and Costa and Kahn (2001) target the prevalence of 'power couples' (where both are college educated) in large cities.

The use of such indirect indicators, however, raises questions. For example, the link between education levels and wages may be direct rather than associated with spillovers (Rauch, 1993). Similarly, skilled labor might be related to labor pooling rather than

knowledge sharing, or to labor supply rather than demand factors (although the actual cause of the spillover may not be as crucial, except for interpretation, as whether the skill intensity is related to agglomeration). Such problems arise because 'workers are the primary vehicle of knowledge spillovers' (Rosenthal and Strange, 2004) and yet worker characteristics can be indicative of other unmeasured factors.

Sometimes researchers thus recognize a variety of factors that may underlie one or more agglomeration drivers. For example, Ellison and Glaeser (1999) use measures of prior innovation in the industry, both manufactured and service inputs, labor specialization proxied by labor productivity, the number of managers per production worker, and educational characteristics as indicators of Marshallian factors affecting agglomeration. They also use measures of natural resources for natural selection and output perishability for transport costs.

Further, such proxies for Marshallian and other agglomeration drivers involve interactions among firms theoretically motivated by the $q(\cdot)$ part of the A_j definition from equation (6.1), but as reflected in the $a(\cdot)$ component of that definition their impacts are expected to differ by temporal, geographical and industrial 'distance'. Empirically representing these dimensions of distance may be even more difficult than finding appropriate proxies for agglomeration causes.[15]

It seems relatively straightforward to identify the temporal or dynamic (lag) nature of agglomeration economies because time is linear (relative to spatial or particularly industrial 'distance'). However, empirical representation of such lags becomes complex because the dynamic effect is cumulative, with impacts occurring over a long period (Glaeser et al., 1992; Henderson et al., 1995). The large literature on the productive contribution of R&D shows that specifying lags for such mechanisms is both imprecise and may have significant impacts on the empirical results (Alston and Pardey, 1996). In the agglomeration economies literature, Henderson (1997) is perhaps the best example of a detailed exploration of lag patterns; he finds that own-industry economies may involve two- to five-year lags.

Geographic distance raises additional problems because it involves defining 'neighbors'. Typically agglomeration economy studies rely on data at one level of spatial aggregation (county, MSA or state), assuming that all productive activity is in the same locality, which precludes recognizing any attenuating effects of geographic distance. However, concentration at different spatial aggregation levels may be recognized, for example by using county-level employment concentration to explain state-level productivity (Ciccone and Hall, 1996). More directly, Rosenthal and Strange (2001) correlate industry characteristics to agglomeration indexes for zip code, county and state levels, and Rosenthal and Strange (2003) measure the distance effects of employment concentration on firm entry using rings of different sizes around an establishment's location. Henderson (2003) similarly considers the productivity effect of employment density in a plant's own county versus neighboring counties.

Creating measures of industry distance is even more difficult. As already noted, measures of diversity or localization economies are typically based on distinguishing productive activity only of 'own' versus 'other' industries or of input suppliers versus demanders. However, some attempts have been made to identify 'co-agglomeration' among industries, such as Dumais et al.'s (2002) representation of connections among two-digit industries.

This measure is constructed as a between-industry version of indexes that are typically used to represent own-industry density. Perhaps the most common of such measures is

the G statistic defined by Audretsch and Feldman (1996) as $G = \Sigma_l(x_l - s_l)^2$, where s_l is location l's share of aggregate employment in an industry (of L locations or regions in the entire area, such as a state or country) and x_l is the location's share of total employment.[16] The difficulty with this index is that it does not take into account the industrial concentration, typically represented by a Herfindahl index of the form $H = \Sigma_j z_j^2$ for the J firms in the industry. The Ellison and Glaeser (1997) index of spatial concentration takes this into account by constructing an index that combines these two measures; $\gamma = [G - (1 - \Sigma_l x_l^2)H]/[(1 - \Sigma_l x_l^2)(1 - H)]$, where γ approaches $G/(1-\Sigma_l x_l^2)$ as the industry approaches perfect competition, but otherwise accommodates excess concentration adapted for industry concentration.

Dumais et al. (2002) extend this to a co-agglomeration 'diversity' index reflecting the correlation in the location choices of plants belonging to different industries. One issue for construction of this measure is defining the R industries that are linked or have co-agglomerative tendencies; typically they are specified as, for example, potential suppliers or demanders (upstream or downstream 'thick markets'; Bartlesman et al., 1994; Ciccone and Hall, 1996). The index is then constructed as: $\gamma^C = [(G/(1 - \Sigma_l x_l^2)) - H - \Sigma_r \gamma_r l_r^2(1 - H_r)]/(1 - \Sigma_r l_r^2)$, where l_r is the r^{th} industry's share of total employment in the R industries, H_r is the Herfindahl index for the r^{th} industry, and $H = \Sigma_r l_r^2 H_r$.

Proxies for agglomeration causes defined as A_j variables can be empirically related to such agglomeration measures to evaluate their contribution to density (Ellison and Glaeser, 1997), or directly to measures of economic performance to determine their productive impacts (Henderson, 2003). The agglomeration measures themselves can also be used as A_j variables and related to economic performance measures (Maurel and Sedillot, 1999). In either case, production theory models that directly represent productivity, such as the production or cost functions in equations (6.2) and (6.3) that include A_j shift factors as arguments, can be used to translate directly the impacts of such factors into agglomeration economy measures.

Model and measures

Once the $\mathbf{A}_j$ (and $\mathbf{x}_j$) arguments of the model are defined, econometric implementation of a production theory model requires specification of the function to be estimated and the agglomeration economy measures to be constructed. It also requires the choice of a functional form to approximate the function, which must be flexible (such as a translog or generalized Leontief) to capture interaction effects, or second-order relationships among the arguments of the function (Paul, 1999).

The use of, say, a production versus cost function,[17] depends in part on the assumptions one wishes to make and the types of measures desired. For example, the production function $y_j = f(\mathbf{x}_j,\mathbf{A}_j)$ represents technological (marginal product) rather than economic (demand or supply) relationships, but it also does not require assumptions such as cost minimization. By contrast, the dual cost function $C_j = g(\mathbf{p}_j,y_j,\mathbf{A}_j)$ allows direct representation of input (such as labor) demands, as well as of agglomeration economies in the form of input cost savings from a change in the A_j variables.

Specifically, using a production function model, agglomeration economy measures reflect the overall productive impact of a change in an $\mathbf{A}_j$ vector component, and the (potentially varying) marginal product impacts for the different inputs of such a change.

These first- and second-order impacts, the latter of which involve cross-effects that are not identified if the functional form is not flexible, may be measured as derivatives in levels or in proportions (elasticities).

For example, the productive impact (marginal product) of, say, R&D activity in a particular location (A_{RD}) may be defined by the derivative $\partial y/\partial A_{RD}$ or the output elasticity or 'share' $\varepsilon_{y,ARD} = \partial \ln y/\partial \ln A_{RD} = \partial y/\partial A_{RD} \cdot (A_{RD}/y)$ (where the firm j subscript is omitted for simplicity). The effect of A_{RD} on the marginal product of labor (x_L) can be measured as the second-order relationship $\partial^2 y/\partial x_L \partial A_{RD}$. Econometric estimation of a production function thus permits the agglomeration productivity effect $\varepsilon_{y,ARD}$ to be distinguished from productivity growth, expressed as $\varepsilon_{y,t} = \partial \ln y/\partial t$ where t is a time counter, and from internal returns to scale, expressed as the sum of the output elasticities for all inputs $\Sigma_n \varepsilon_{y,n} = \Sigma_n \partial \ln y/\partial \ln x_n$. It also allows the input-specific agglomeration impacts to be distinguished from biased technical change and input substitutability, captured by $\partial^2 y/\partial x_L \partial t$ and $\partial^2 y/\partial x_L \partial x_n$ $(n \neq L)$.

Agglomeration economy measures for a cost function are similar, but more explicitly represent cost changes (which are typically how agglomeration economies are defined), input demands, and internal versus external scale economies. That is, the overall productivity effect of, say, greater density of R&D activity in a particular location (A_{RD}) may be measured in terms of the first-order cost effect $\partial C/\partial A_{RD}$ or the (proportional) elasticity $\varepsilon_{C,ARD} = \partial \ln C/\partial \ln A_{RD} = \partial C/\partial A_{RD} \cdot (A_{RD}/C)$. This cost-saving effect is therefore explicitly an 'economy' associated with more A_{RD}, or a downward shift of the cost curve in locations with more A_{RD}. The underlying input demand impacts can be separately identified based on Shephard's lemma; $x_n = \partial C/\partial p_n$ for input n (say, labor, L), so the labor demand effect of more A_{RD} may be measured as the second derivative $\partial^2 C/\partial p_L \partial A_{RD} = \partial x_L/\partial A_{RD}$ or the second-order elasticity $\varepsilon_{xL,ARD} = \partial^2 \ln C/\partial \ln p_L \partial \ln A_{RD} = \partial \ln x_L/\partial \ln A_{RD}$.

Again, these economies from external spillovers or inputs represented as A_j variables can be econometrically distinguished in a full cost function model from temporal productivity growth, $\varepsilon_{C,t} = \partial \ln C/\partial t$, as well as from internal scale economies, $\varepsilon_{C,y} = \partial \ln C/\partial \ln y$ (for one output or $\Sigma_m \varepsilon_{C,ym} = \Sigma_m \partial \ln C/\partial \ln y_m$ for multiple outputs, where the cross-terms between the outputs reflect scope economies). Input-specific technical change and substitution effects can also be distinguished separately from the agglomeration effects if a flexible functional form is used.[18]

Econometric issues

The next step is to specify the econometric model. The primary econometric issues raised in the empirical literature on agglomeration economies involve measurement error from omitted variables and endogeneity or causality (Eberts and McMillen, 1999). Spatial autocorrelation or spatial error term 'lags' also become an issue, however, when production and performance depend on location. Because spatial autocorrelation can arise from omitted variables that vary across geographic space, an econometric model that accounts for spatial autocorrelation can rectify some measurement error issues.

The issue of omitted variables has arisen in studies such as Sveikauskas (1975), which lacked capital data. Moomaw (1981) showed that this could significantly upward bias the coefficient estimates because capital intensity might be expected to be greater in larger cities, whereas a lack of land data could have the opposite effect since land availability is limited in cities.

Endogeneity or causality issues arise if, for example, higher wages in cities could be due to agglomeration causes or, conversely, urbanization could occur because more productive workers are attracted to cities. Such problems call for the use of econometric techniques such as instrumental variables, although appropriate instruments (that are exogenous to the dependent variable, such as productivity, but correlated with the agglomeration measures) are often difficult to obtain, and the results may be sensitive to the (arbitrary) choice of instruments. For example, Henderson (2003) uses (exogenous) local environment indicators or lagged values of marketplace characteristics such as types of local employment as instruments, but finds them to be weak instruments. He thus instead relies on fixed effects for locality and time that are meant to reflect unobserved factors affecting firm location. Other studies have used lagged values of the agglomeration measures as instruments (Glaeser et al., 1992; Henderson et al., 1995; Rosenthal and Strange, 2003).

Spatial autocorrelation has received much less attention in the literature than these other econometric issues, although it is important to recognize if omitted variables vary spatially. Spatial autocorrelation (Anselin, 1988) is the spatial analogue to a temporal autocorrelation or autoregressive adjustment; it involves dependence of the error terms for a particular location on the weighted average of errors for nearby locations. Like temporal autocorrelation, the lag structure that might be at work must be specified before correcting for it. However, in the spatial context it is not straightforward to define the 'neighbors' that might exhibit such dependence.

More specifically, spatial linkages are accommodated in the stochastic structure via 'lags' for geographic location at any one point in time. If there is only interaction between two firms in nearby locations the spatial autoregressive (SAR) adaptation takes the form $u_{k,t} = \rho u_{j,t} + \varepsilon_{k,t}$ where $u_{j,t}$ is the (unadjusted) error term for firm j at time t and $\varepsilon_{k,t}$ is an independently, identically distributed error.[19] If activities in multiple locations affect firm (or region) k's error term, it instead becomes a weighted sum of the errors for other firms (or regions): $u_{k,t} = \rho\Sigma_j w_{k,j} u_{j,t} + \varepsilon_{k,t}$ (Cohen and Paul, 2004), where $w_{k,j}$ is the weight that spatial unit j has on unit k's error term. Such a model allows consideration of whether spatial lags in the econometric specification are significant by testing the significance of the ρ parameter.

One challenge for applying the SAR model therefore is specifying which 'neighbors' exhibit stochastic spillovers and how to weight their impacts. For example, in Cohen and Paul (2004) spillovers from transportation infrastructure are assumed to occur across states. The interrelated states are thus defined as those with a common boundary, and the $w_{k,j}$ for each equation defined as giving all neighboring states equal weight and all other states zero weight.

If there are additional layers of dependencies, so higher-order autocorrelation spills over but possibly attenuates over distance, the specification becomes further complicated (Cohen and Paul, 2007). In such a situation with two bands of neighbors, the error term would be written as $u_{k,t} = \rho_1\Sigma_j w_{1,k,j} u_{j,t} + \rho_2\Sigma_i w_{2,k,t} u_{i,t} + \varepsilon_{k,t}$, where ρ_1 and ρ_2 are the impacts of the average of the first-order (j) and second-order (i) neighbors' errors on the particular (k) unit's error, and $w_{1,k,j}$ and $w_{2,k,i}$ are the weights for the first- and second-order neighbors on a particular unit's errors.

6.5 Specific applications based on production theory models

To illustrate the potential of production theory models – in particular cost-based models – for evaluation of agglomeration economies, we will briefly summarize three applications

that address various issues raised in this chapter. The first involves the potential for labor pooling, in the context of hospitals (Cohen and Paul, forthcoming).

The literature on the cost efficiency of hospital services includes some applications that explore (internal) scale and scope economies, due to important questions about health cost efficiency, but has rarely addressed agglomeration economies. Measures of scale and scope economies, which can be directly computed from cost function models as alluded to above, allow consideration of enhanced cost efficiency from size and diversification. Cohen and Paul (forthcoming) construct and assess such measures, but also recognize that geographical proximity of hospitals permits both labor pooling and knowledge (expertise and capital) sharing, so hospital clustering may be cost-saving (efficiency-enhancing).

The limited literature on this aspect of cost efficiency for hospitals includes O hUallachain and Satterthwaite (1992) and Bates and Santerre (2005), who use metropolitan data to represent agglomeration economies by the number of hospitals and scale economies by hospital size, based on one aggregate measure of hospital output. Cohen and Paul (forthcoming) instead use hospital-level data and a cost function model that includes as arguments two outputs (y_{mj}) to facilitate evaluation of substitution between outpatient and inpatient services, nine labor types (x_{nj}) that distinguish treatment types, and capacity and Medicare[20] percentages as hospital-specific factors. They also include an agglomeration (A_j) variable to represent the proximity of other hospitals[21] and a correction for spatial autocorrelation to identify spatial linkages in the econometric specification. For econometric implementation they use a flexible (generalized Leontief) functional form for the cost function, and estimate a system comprised of this function and the input demand functions.

Cohen and Paul (forthcoming) find significant agglomeration economies as well as spatial autocorrelation; knowledge sharing through proximity to other hospitals measured as $\varepsilon_{C,Aj}$ appears to be cost-saving and the spatial ρs are significant. In turn, second-order elasticities reveal that more outpatient visits relative to inpatient days reduces the value of hospital clustering, and that labor cost impacts vary by treatment center (but clustering reduces costs for all services except psychiatric and 'other inpatient' services). Significant scale economies for outpatient visits are also found, as is some evidence of scope economies via output and treatment center complementarities.

Cohen and Paul (2004) also evaluate the cost impact of an external variable – in this case the availability of transportation infrastructure. They investigate production cost savings from substitution of public capital (infrastructure) for private inputs, using state-level US manufacturing data on prices and quantities of aggregate output and (capital, production and non-production labor, and materials) inputs. Many studies have used production or cost function models to address such issues, with infrastructure included as an A_j variable (Boarnet, 1998; Conrad and Seitz, 1992; Morrison and Schwartz, 1996). Spatial spillovers from public capital investment in geographically linked areas has received less attention, although some studies have raised this possibility (Kelejian and Robinson, 1997; Holtz-Eakin and Schwartz, 1995; Boarnet, 1998).

Cohen and Paul (2004) estimate a flexible cost function model using data on stocks of public highway infrastructure in both the own and neighboring states as A_j variables. The model thus distinguishes intra- and interstate impacts of public infrastructure stock levels and their interdependencies, or geographic layers of spillovers. The econometric model is

also adapted to measure the extent and significance of spatial spillovers using SAR techniques. They find significant cost savings from intra-state public infrastructure investment that is both enhanced and augmented by cross-state spillovers, as well as increasing intra- and interstate public capital impacts over time.

A similar model is used by Cohen and Paul (2007) to evaluate the impacts on the shadow (asset) values of private capital stocks from both highway and airport infrastructure in a state. The model allows for the possibility of intermodal interdependencies between highways and airports (more airports could, for example, imply more congestion on highways, reducing the productive contribution of this public infrastructure). It also allows not only for one level but also for layers of spatial autocorrelation across states, with first-order 'neighboring' states defined as those with common boundaries and second-order states defined as neighbors of neighbors. Cohen and Paul (2007) find significant impacts on asset shadow values and input composition from both kinds of public infrastructure investment that have a ripple effect; at least one and as many as three spatial error lags are significant.

Finally, Cohen and Paul (2005) use such a framework to focus on the industrial scope of agglomeration economy spillovers and resulting location implications. They model and measure thick market effects for food processing industries from proximity to density of own-industry production and to demanders (consumers) and suppliers (primary agricultural production). They find significant average cost savings for food processors from locating close to own-industry markets, suppliers and demanders (thick market effects), but higher production costs associated with greater within-state agricultural intensity (thin market effects). By contrast, marginal costs are higher in more urban and lower in more rural areas. They also find that geographic concentration patterns of US food processors, or location decisions, seem to be motivated by such cost considerations.

6.6 Concluding remarks

In this chapter we have focused on the potential of production theory models to model and measure agglomeration economies. We have emphasized the challenges involved in defining and measuring the arguments of production or cost function models that might be used for such purposes – in particular the agglomeration 'causes' modeled as 'A_j' variables that act as shift factors for the production or cost frontier. However, we have also shown the potential of such models for providing empirical insights about both the overall productive 'effects' of such factors (in terms of productivity or costs), and the underlying input-specific impacts and other interactions among functional arguments. Such models also provide the basis for evaluating location decisions of firms and resulting regional growth that are motivated by the productive effects of clustering. We anticipate seeing the empirical literature in this area expand in this direction.

Notes

1. This may also be true for individuals, although, as noted above, our focus in this chapter is on production.
2. This may in turn imply 'lumpy' and expensive inputs that may not be fully utilized by firms in isolation.
3. This notion of input sharing also implies vertical disintegration so input suppliers are more specialized.
4. Other possibilities focused more on individuals rather than firms, such as economies of consumption that imply regional growth because people have amenities such as good restaurants close by (Glaeser et al., 2001) or that job seeking for 'power couples' (Costa and Kahn, 2001) attracts them to cities have also been suggested, although due to our primary focus on firms we will not discuss these further.

5. Such an effect may also involve trade, for example if trade liberalization lessens the importance of home markets.
6. The mechanisms through which these effects occur have also been explored, such as whether the productivity impacts of density are due to local competition forcing innovation (Porter, 1990) or to local culture (Saxenian, 1994; Rosenthal and Strange, 2006), although that is not the focus of this chapter,
7. However, such studies are typically carried out at higher levels of spatial aggregation such as a city, county or state.
8. Note that density could also impose costs due to, say, congestion (Maggioni, 2002; Ottaviano and Thisse, 2004). Although we will refer to agglomeration economies as positive externalities in this chapter, it is therefore possible that these negative externalities could counteract positive effects, resulting in negative productive or cost impacts of firm concentration or density.
9. Alternatively they may be related to concentrations of people, although firms are our focus here.
10. Labor supply may also be involved; for example an urban area may attract people such as 'power couples' (Costa and Kahn, 2001) or those more amenable to the 'urban rat race' (Rosenthal and Strange, 2008). However, due to our focus on firm behavior and performance we will not explore this literature.
11. See, for example, Shefer (1973), Sveikauskas (1975), Segal (1976), Fogarty and Garofalo (1978), Moomaw (1981) and Tabuchi (1986).
12. Multiple outputs also may be represented in cost function models, as discussed below in the context of hospitals, which facilitates separate consideration of scope economies, although the literature tends to be based on aggregated output.
13. See Rosenthal and Strange (2004), pp. 2134–5 for further discussion of these issues.
14. Glaeser et al. (1992) evaluate the growth of the six largest industries in a particular area and its dependence on the share of employment of the 7th through 12th largest industries to represent a diverse industrial base. Henderson et al. (1995) use a Herfindahl index of representing employment diversity relative to a uniform distribution.
15. The impacts of these factors would also be expected to differ by industry and country. Such differences are typically explored by analyzing agglomeration effects for data that is sufficiently disaggregated along these dimensions to identify the differences (Nakamura, 1985; Henderson, 1986; Henderson et al., 1995, for example, by industry and Ciccone, 2002, by country).
16. G is thus equal to zero when an industry and total employment are identically spatially distributed and to one if the industry is fully concentrated in one location.
17. Other functions such as a profit or revenue function may be useful for some purposes, but in this literature production and cost functions are the most common. For further discussion of the choice of functional representation as well as functional form see Paul (1999).
18. Note that such internal scale economies involve movements along the long-run cost curve by contrast to a shift in the curve from external economies, which raises two associated issues: (1) whether economies are external or internal depends on the level of aggregation since, for example for state-level data, externalities among firms or even counties will be internalized; and (2) for some applications (and data) it may be important to distinguish short- from long-run economies, which involves recognizing fixed inputs as x_n levels in the (variable) cost function and imputing shadow values (Morrison and Siegel, 1999).
19. Note that this adaptation is somewhat analogous to an AR(1) adjustment (a first order auto regressive process standard in the time series literature). However, the analogy is not perfect because the ordering of the observations in spatial models is not important, whereas in time series models preserving the ordering of the observations is crucial.
20. Medicare is a government-sponsored health insurance program in the US for adults above age 65.
21. This variable is defined as a spatially weighted average of other hospitals' labor force (full-time equivalents).

References

Alston, J.M. and P.G. Pardey (1996), *Making Science Pay: Economics of Agricultural R&D Policy*, Washington, DC: American Enterprise Institute for Public Policy.

Anselin, L. (1988), *Spatial Econometrics: Methods and Models*, Boston, MA: Kluwer Academic Publishers.

Audretsch, D.B. (1998), 'Agglomeration and the location of innovative activity', *Oxford Review of Economic Policy*, **14** (2), 18–29.

Audretsch, D.B. and M.P. Feldman (1996), 'R&D spillovers and the geography of innovation and production', *American Economic Review*, **86** (3), 630–40.

Baptista, R. (1997), 'An Empirical Study of Innovation, Entry and Diffusion in Industrial Clusters', PhD dissertation, University of London, London Business School.

Barrios, S., L. Bertinelli, E. Strobl and A.C. Teixeira (2005), 'The dynamics of agglomeration: evidence from Ireland and Portugal', *Journal of Urban Ec onomics*, **57**, 170–88.

Bartlesman, E.J., R.J. Caballero and R.K. Lyons (1994), 'Customer and supplier-driven externalities', *American Economic Review*, **84** (4), 1075–84.
Bates, L.J. and R.E. Santerre (2005), 'Do agglomeration economies exist in the hospital services industry?', *Eastern Economic Journal*, **31** (4), 617–28.
Baumgardner, J.F. (1988), 'Physicians' services and the division of labor across local markets', *Journal of Political Economy*, **96** (5), 948–82.
Boarnet, Marlon (1998), 'Spillovers and the locational effects of public infrastructure', *Journal of Regional Science*, **38**, 381–400.
Carlton, D.W. (1983), 'The location and employment choices of new firms: an econometric model with discrete and continuous endogenous variables', *Review of Economics and Statistics*, **65**, 440–49.
Chapple, W., R. Harris and C.J. Morrison Paul (2006), 'The cost implications of waste reduction: factor demand, competitiveness and policy implications', *Journal of Productivity Analysis*, **26** (3), 245–58.
Ciccone, A. (2002). 'Agglomeration effects in Europe', *European Economic Review*, **46**, 213–27.
Ciccone, A. and R.E. Hall (1996), 'Productivity and the density of economic activity', *American Economic Review*, **86**, 54–70.
Cohen, J.P. and C.J. Morrison Paul (2004), 'Public infrastructure investment, costs, and inter-state spatial spillovers in US manufacturing: 1982–96', *Review of Economics and Statistics*, **86** (2), 551–60.
Cohen, J.P. and C.J. Morrison Paul (2005), 'Agglomeration economies and industry location decisions: the impacts of spatial and industrial spillovers', *Regional Science and Urban Economics*, **35** (3), 215–37.
Cohen, J.P. and C.J. Morrison Paul (2007), 'Higher order spatial autocorrelation in a system of equations: the impacts of transportation infrastructure on capital asset values', *Journal of Regional Science*, **47** (13), 457–78.
Cohen, J.P. and C.J. Morrison Paul (forthcoming), 'Scale, scope, and agglomeration economies in hospital services', *Regional Science and Urban Economics*.
Conrad, K. and H. Seitz (1992), 'The "Public Capital Hypothesis": the case of Germany', *Recherches Economiques de Louvain*, **58** (3–4), 309–27.
Costa, D.L. and M.E. Kahn (2001), 'Power couples', *Quarterly Journal of Economics*, **116**, 1287–1315.
Davis, D.R. and D.E. Weinstein (1999), 'Economic geography and regional production structure: an empirical investigation', *European Economic Review*, **43**, 379–407.
Dekle, R. and J. Eaton (1999), 'Agglomeration and land rents, evidence from the prefectures', *Journal of Urban Economics*, **46**, 200–214.
Diamond, C.A. and C.J. Simon (1990), 'Industrial specialization and the returns to labor', *Journal of Labor Economics*, **8** (2), 175–201.
Dumais, G., G. Ellison and E. Glaeser (2002), 'Geographic concentration as a dynamic process', *Review of Economics and Statistics*, **84** (2), 193–204.
Duranton, G. and D. Puga (2001), 'Nursery cities: urban diversity, process innovation, and the life-cycle of products', *American Economic Review*, **91** (5), 1454–77.
Eberts, R.W. and D.P. McMillen (1999), 'Agglomeration economies and urban public infrastructure', in P. Cheshire and E.S. Mills (eds), *Handbook of Regional and Urban Economics*, Vol. 3, New York: North-Holland, pp. 1455–95.
Ellison, G. and E.L. Glaeser (1997), 'Geographic concentration in US manufacturing industries: a dartboard approach', *Journal of Political Economy*, **105**, 889–927.
Ellison, G. and E. Glaeser (1999), 'The geographic concentration of an industry: does natural advantage explain agglomeration?', A*merican Economic Association Papers and Proceedings*, **89**, 311–16.
Fogarty, M.S. and G.A. Garofalo (1978), 'Urban spatial structure and productivity growth in the manufacturing sector of cities', *Journal of Urban Economics*, **23**, 60–70.
Fuchs, V. (1962), *Changes in the Location of Manufacturing in the US since 1929*, New Haven, CT: Yale University Press.
Glaeser, E.L. and D.C. Mare (2001), 'Cities and skills', *Journal of Labor Economics*, **19** (2), 316–42.
Glaeser, E., J. Kolko and A. Saiz (2001), 'Consumer city', *Journal of Economic Geography*, **1**, 27–50.
Glaeser, E.L., H.D. Kallal, J.A. Scheinkman and A. Shleifer (1992), 'Growth in cities', *Journal of Political Economy*, **100**, 1126–52.
Griliches, Z. (1979), 'Issues in assessing the contribution of R&D to productivity growth', *Bell Journal of Economics*, **10**, 92–116.
Henderson, J.V. (1986), 'Efficiency of resource usage and city size', *Journal of Urban Economics*, **19**, 47–70.
Henderson, J.V. (1997), 'Externalities and industrial development', *Journal of Urban Economics*, **42**, 449–70.
Henderson, J.V. (2003), 'Marshall's scale economies', *Journal of Urban Economics*, **43**, 1–28.
Henderson, J.V., A. Kuncoro and M. Turner (1995), 'Industrial development in cities', *Journal of Political Economy*, **103**, 1067–85.
Holmes, T.J. (1999), 'Localization of industry and vertical disintegration', *Review of Economics and Statistics*, **81** (2), 314–25.

Holmes, T.J. and J.J. Stevens (2002), 'Geographic concentration and establishment scale', *Review of Economics and Statistics*, **84** (4), 682–91.
Holtz-Eakin, D. and A.E. Schwartz (1995), 'Spatial productivity spillovers from public infrastructure: evidence from state highways', *International Tax and Public Finance*, **2**, 459–68.
Jaffe, A.G., M. Trajtenberg and R. Henderson (1993), 'Geographic localization of knowledge spillovers as evidenced by patent citations', *Quarterly Journal of Economics*, **108**, 577–98.
Kelejian, H. and D. Robinson (1997), 'Infrastructure productivity estimation and its underlying econometric specifications: a sensitivity analysis', *Papers in Regional Science*, **76** (1), 115–31.
Kim, S. (1995), 'Expansion of markets and the geographic distribution of economic activities: the trends in US regional manufacturing structure, 1860–1987', *Quarterly Journal of Economics*, **110** (4), 881–908.
Kim, S. (1999), 'Regions, resources and economic geography: sources of US regional comparative advantage, 1880–1987', *Regional Science and Urban Economics*, **29**, 1–32.
Krugman, P. (1980), 'Scale economies, product differentiation, and the pattern of trade', *American Economic Review*, **70**, 950–59.
Krugman, P. (1991), *Geography and Trade*, Cambridge, MA: MIT Press.
Maggioni, M. (2002), *Clustering Dynamics and the Location of High-Tech Firms*, Heidelberg and New York: Physica-Verlag.
Marshall, A. (1920), *Principles of Economics*, London: Macmillan.
Maurel, F. and B. Sedillot (1999), 'A measure of the geographic concentration in French manufacturing industries', *Regional Science and Urban Economics*, **5**, 575–604.
McDonald, J. and D. McMillen (2007), *Urban Economics and Real Estate: Theory and Policy*, Malden, MA: Blackwell Publishing.
Moomaw, R.L. (1981), 'Productivity and city size: a critique of the evidence', *Quarterly Journal of Economics*, **96**, 675–88.
Moreno, R., E. Lopez-Bazo, E. Vaya and M. Artis (2004), 'External effects and cost of production', in L. Anselin, R. Florax and S. Rey (eds), *Advances in Spatial Econometrics: Methodology, Tools, and Applications*, Berlin/Heidelberg/New York: Springer Press, pp. 297–317.
Moretti, E. (2004), 'Human capital externalities in cities', in J.V. Henderson and J.F. Thisse (eds), *Handbook of Regional and Urban Economics*, Vol. 4, Amsterdam: Elsevier Press, pp. 2243–91.
Morrison, C.J. and A.E. Schwartz (1996), 'State infrastructure and productive performance', *American Economic Review*, **86** (5), 1095–1111.
Morrison, C.J. and D. Siegel (1998), 'Knowledge capital and cost structure in the US food and fiber industries', *American Journal of Agricultural Economics*, **80** (1), 30–45.
Morrison, C.J. and D. Siegel (1999), 'Scale economies and industry agglomeration externalities: a dynamic cost function approach', *American Economic Review*, **89** (1), 272–90.
Nakamura, R. (1985), 'Agglomeration economies in urban manufacturing industries: a case of Japanese cities', *Journal of Urban Economics*, **17**, 108–24.
O'Sullivan, A. (2007), *Urban Economics*, 6th edn, New York, NY: McGraw-Hill.
Ohlin, B. (1933), *Interregional and International Trade*, Cambridge, MA: Harvard University Press.
O hUallachain, B. and M.A. Satterthwaite (1992), 'Sectoral growth patterns at the metropolitan level, an evaluation of economic development incentives', *Journal of Urban Economics*, **31**, 25–58.
Ottaviano, G. and J.F. Thisse (2004), 'Agglomeration and economic geography', in J.V. Henderson and J.F. Thisse (eds), *Handbook of Regional and Urban Economics*, Vol. 4, Amsterdam: Elsevier Press, pp. 2563–2608.
Paul, C.J. Morrison (1999), *Cost Structure and the Measurement of Economic Performance: Productivity, Utilization, Cost Economies and Related Performance Indicators*, Boston, MA: Kluwer Academic Press.
Porter, M. (1990), *The Competitive Advantage of Nations*, New York: Free Press.
Rauch, J. (1993), 'Productivity gains from geographic concentration of human capital: evidence from the cities', *Journal of Urban Economics*, **34**, 380–400.
Roback, J. (1982), 'Wages, rents, and the quality of life', *Journal of Political Economy*, **90**, 1257–78.
Rosen, S. (1979), 'Wage-based indexes of urban quality of life', in P. Mieszkowski and M. Straszheim (eds), *Current Issues in Urban Economics*, Baltimore, MD: Johns Hopkins University Press, pp. 391–429.
Rosenthal, S.S. and W.C. Strange (2001), 'The determinants of agglomeration', *Journal of Urban Economics*, **50**, 191–229.
Rosenthal, S.S. and W.C. Strange (2003), 'Geography, industrial organization, and agglomeration', *Review of Economics and Statistics*, **85** (2), 377–93.
Rosenthal, S.S. and W.C. Strange (2004), 'Evidence on the nature and sources of agglomeration economies', in J.V. Henderson and J.F. Thisse (eds), *Handbook of Regional and Urban Economics*, Vol. 4, Amsterdam: Elsevier Press, pp. 2019–2171.
Rosenthal, S.S. and W.C. Strange (2006), 'The micro-empirics of agglomeration economies', in Daniel P. McMillen and Richard Arnott (eds), *A Companion to Urban Economics*. Malden, MA: Blackwell Press, pp. 7–23.

Rosenthal, S.S. and W.C. Strange (2008), 'Agglomeration and hours worked', *Review of Economics and Statistics*, **90** (1), 105–18.
Rosenthal, S.S. and W.C. Strange (forthcoming), 'The attenuation of human capital spillovers', *Journal of Urban Economics*.
Saxenian, A. (1994), *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*, Cambridge, MA: Harvard University Press.
Segal, D. (1976), 'Are there returns to scale in city size?', *Review of Economics and Statistics*, **58**, 339–50.
Shefer, D. (1973), 'Localization economies in SMSAs: a production function analysis', *Journal of Regional Science*, **13**, 55–64.
Sveikauskas, L.A. (1975), 'The productivity of cities', *Quarterly Journal of Economics*, **89**, 393–413.
Tabuchi, T. (1986), 'Urban agglomeration, capital augmenting technology, and labor market equilibrium', *Journal of Urban Economics*, **20**, 211–28.
Von Hipple, E. (1994), 'Sticky information and the locus of problem solving: implications for innovation', *Management Science*, **40**, 429–39.
Wheaton, W.C. and M.J. Lewis (2002), 'Urban wages and labor market agglomeration', *Journal of Urban Economics*, **51**, 542–62.

7 Territorial capital and regional development

Roberto Camagni

7.1 The resurgence of supply-oriented approaches

Looking at the recent evolution of theoretical regional economics, we may argue that, in the long term, supply-oriented approaches have outperformed strictly demand-oriented ones, of a Keynesian nature, in the interpretation of regional development processes.

In fact, on the one hand, regional internal demand is not relevant, even in the short run, to drive regional growth, given the huge interregional integration and ever-increasing international division of labour. On the other hand, national demand growth is certainly more relevant to internal regional performances, but it is so on a 'on-average' basis: single regions may outperform (or underperform) the national average at the expense (in favour of) other regions,[1] either because of a more appropriate (poorer) sectoral mix or because of a favourable (unfavourable) competitive differential.

International demand growth, too, in particular as regards specific productions, may be highly favourable to the development of specific regions specialised in high-growth demand sectors. But this relationship may probably work well in a first approximation and in the short run; in a more precise and longer-term perspective, there is no necessary reason why different regions should benefit equally from the (aggregate or sectoral) expansion of international trade. Textiles, shipbuilding or car production were for long considered slow-growing industries, but this fact did not prevent the emergence of regional and national success stories such as, respectively, Tuscany, Korea or Japan, areas that proved able to acquire rapidly increasing shares of an even stagnant international market.

From an *ex ante* and logical point of view, it is exactly this regional differential growth capability that must be interpreted, and possibly forecasted, on the basis of supply-side elements.

Integrated demand–supply approaches based on complex feedback effects between demand-driven shoves and increasing returns effects have for long shown good explanatory capacity, especially when strong cumulative effects, either virtuous or vicious, have been widely apparent and pervasively affecting broad typologies of winner and loser regions.

Today, a more selective pattern of regional growth is emerging. It differentiates among the development paths of single regions and produces a varied mosaic of development stories. This phenomenon calls for more stringent and selective interpretations of the different regional development paths. Perhaps scholars themselves are becoming more demanding in terms of the specific interpretation of region-specific growth paths, and more sensitive to the consequent need to build tailor-made growth strategies for each territory.

This awareness is today strengthened by a new crucial theoretical argument: in a context of globalisation and the creation of broad single-currency areas, regions (and also nations) must closely concern themselves with the competitiveness of their production systems because no spontaneous or automatic adjustment mechanism is still at work

to counterbalance a lack (or an insufficient growth rate) of productivity. Currency devaluation is no longer viable (by definition in the case of regions; by international monetary agreements in the case of countries), nor are international monetary agreements; and wage–price flexibility is not sufficient or rapid enough to restore equilibrium once it has been perturbed, mainly because wages and prices are not determined on a regional base. In terms of international and interregional trade theory, regions do not compete with each other on the basis of a Ricardian 'comparative advantage' principle – which guarantees each region a role in the international division of labour[2] – but rather on a Smithian 'absolute advantage' principle, similar in nature to Porter's concept of 'competitive advantage' (Camagni, 2002).

Therefore, regional and local governments must address the issue of the competitiveness and attractiveness of external firms. Definition of possible growth strategies for each region, city or territory must necessarily rely on local assets and potentials and their full – and wise – exploitation: in short, on what is increasingly called 'territorial capital'.

7.2 Towards a cognitive approach to regional development: the concept of territorial capital

Does the above signify that, in terms of interpretive theoretical tools, we are back with traditional, supply-side neoclassical models? In a sense, yes, as local competitiveness cannot but be linked to local supply conditions. But these supply conditions must perforce refer to factors completely different from the traditional ones – namely capital and labour, local resources, and infrastructure endowment. The huge theoretical heritage of the endogenous development literature – industrial districts, *milieux innovateurs*, production clusters – has long directed regional scholars' attention to intangible, atmosphere-type, local synergy and governance factors: what after 1990 were reinterpreted in the form of social capital (Putnam, 1993), relational capital (Camagni, 1999; Camagni and Capello, 2002) or, in a slightly different context, as knowledge assets (Foray, 2000; Storper, 2003; Camagni, 2004).

The shift is not at all merely terminological: a cognitive approach is increasingly superseding the traditional functional approach to show that cause–effect, deterministic relationships should give way to other kinds of complex, inter-subjective relationships which impinge on the way economic agents perceive economic reality, are receptive to external stimuli, can react creatively, and are able to cooperate and work synergetically. Local competitiveness is interpreted as residing in local trust and a sense of belonging rather than in pure availability of capital; in creativity rather than in the pure presence of skilled labour; in connectivity and relationality more than in pure accessibility; in local identity besides local efficiency and quality of life.

The theoretical elements that support the new methodological approach may be found in the following:

- The theory of bounded rationality and decision-making under conditions of uncertainty, from the seminal contributions of Malmgren and Simon (Malmgren, 1961; Simon, 1972) to their application to industrial innovation (Nelson and Winter, 1982; Dosi, 1982).
- The institutional approach to economic theory based on a 'theory of contracts' which emphasizes the importance of rules and behavioural codes, and of institutions that 'embed transactions in more protective governance structures'

(Williamson, 2002, p. 439), reducing conflicts and allowing mutual advantages to be gained from exchange.

- The cognitive approach to district economies and synergies, which comprises the Italian school (Becattini, 1990), the French 'proximity' approach (Gilly and Torre, 2000), the GREMI[3] approach to local innovative environments (Camagni, 1991; Camagni and Maillat, 2006), and Michael Storper's concept of 'untraded interdependencies' (Storper, 1995). The GREMI group conceives proximity space or the local 'milieu' as an uncertainty-reducing operator which works through the socialised transcoding of information, enhancing cooperation, and the supply of the cognitive substrate – represented mainly by the local labour market – in which processes of collective learning are embedded (Camagni, 1991; Capello, 2001).

All the above elements – which add to, and do not substitute for, more traditional, material and functional approaches – may be encompassed and summarized by a concept that, strangely enough, has only recently made its appearance, and has done so outside a strictly scientific context: the concept of territorial capital. This was first proposed in a regional policy context by the Organisation for Economic Co-operation and Development (OECD) in its *Territorial Outlook* (OECD, 2001), and it has been recently reiterated by DG Regio of the Commission of the European Union:

> Each Region has a specific 'territorial capital' that is distinct from that of other areas and generates a higher return for specific kinds of investments than for others, since these are better suited to the area and use its assets and potential more effectively. Territorial development policies (policies with a territorial approach to development) should first and foremost help areas to develop their territorial capital. (European Commission, 2005, p. 1)

As is widely apparent from current research work, 'territory' is a better term than (abstract) 'space' when referring to the following elements:

- A system of localised externalities, both pecuniary (where their advantages are appropriated through market transactions) and technological (when advantages are exploited by simple proximity to the source).
- A system of localised production activities, traditions, skills and know-hows.
- A system of localised proximity relationships which constitute a 'capital' – of a social psychological and political nature – in that they enhance the static and dynamic productivity of local factors.
- A system of cultural elements and values which attribute sense and meaning to local practices and structures and define local identities; they acquire an economic value whenever they can be transformed into marketable products – goods, services and assets – or they boost the internal capacity to exploit local potentials.
- A system of rules and practices defining a local governance model.

Accordingly, the OECD has rightly drawn up a long, sometimes plethoric but well-structured list of factors acting as the determinants of territorial capital, and which range from traditional material assets to more recent immaterial ones.

> These factors may include the area's geographical location, size, factor of production endowment, climate, traditions, natural resources, quality of life or the agglomeration economies

> provided by its cities, but may also include its business incubators and industrial districts or other business networks that reduce transaction costs. Other factors may be 'untraded interdependencies' such as understandings, customs and informal rules that enable economic actors to work together under conditions of uncertainty, or the solidarity, mutual assistance and co-opting of ideas that often develop in clusters of small and medium-sized enterprises working in the same sector (social capital). Lastly, according to Marshall, there is an intangible factor, 'something in the air', called the 'environment' and which is the outcome of a combination of institutions, rules, practices, producers, researchers and policy makers that make a certain creativity and innovation possible. (OECD, 2001, p. 15)

Given these premises, the new concept of territorial capital deserves closer inspection, and mainly in regard to its components and economic meaning. On the one hand, it is clear that some items in the above list belong to the same abstract factor class and differ only in terms of the theoretical approach of their proponents, while some others are lacking. On the other hand, whether the notion of 'capital' can be applied to many of these factors is questionable, because they do not imply an investment, an asset requiring a remuneration, or a production factor expressed in quantitative terms. Nevertheless, at least for the 'material' items comprised in the definition of territorial capital, their use in a 'quasi-production function' is widely justified, following the long tradition launched, *inter alia*, by Biehl (1986) with physical infrastructure, Aschauer (1989) with social public capital and Capello (1994) with information links.

The next section proposes a possible theoretical taxonomy.

7.3 Territorial capital: a theoretical taxonomy

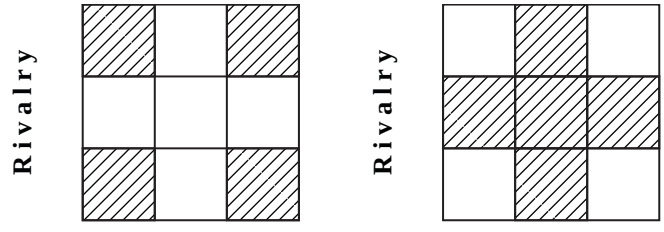
A three-by-three matrix, both theoretically sound and relatively exhaustive, can be proposed to classify all potential sources of territorial capital. It is built upon two main dimensions:

- rivalry: public goods, private goods and an intermediate class of club goods and impure public goods; and
- materiality: tangible goods, intangible goods and an intermediate class of mixed, hard–soft goods.

The four extreme classes – high and low rivalry, tangible and intangible goods – represent by and large the classes of sources of territorial capital usually cited by regional policy schemes. They can be called the 'traditional square'. On the other hand, the four intermediate classes represent more interesting and innovative elements on which new attention should be focused; they can be called the 'innovative cross' (Figure 7.1).

These latter components comprise, on the materiality axis, mixed goods characterised by an integration of hard and soft elements, material goods and services which indicate a capacity to translate virtual and intangible elements into effective action, cooperation, public–private partnership, supply of services; a capacity, that is, to convert potential relationality into effective relationality and linkages among economic agents. On the rivalry axis there is an intermediate class of goods encompassing two different relevant cases:

- impure public goods in which, as in pure public goods, excludability is low, but rivalry is higher because of increasing congestion and scarcity, for example. In this



The 'traditional square' The 'innovative cross'

Figure 7.1 Traditional and innovative factors of territorial capital

case, rivalry may also take the form of interest conflicts among different types of users or between the class of generic (and respectful) users and some specific free-riders whose action may endanger the consistency of the public territorial goods;

- club goods, where the opposite condition holds, namely high excludability (with respect to non-members) and low rivalry.

A third intermediate class, likened here to the category of private goods, can be represented by 'toll goods': a typology of public goods whose use, because it is excludable, is subject to a toll levied by the public administration or by a concessionaire. The closer the price paid is to the production and maintenance cost, the less these public goods are distinguishable from ordinary private goods.

In all these intermediate cases, a crucial control function must be performed by public authorities in order to keep the potential benefit to the local community high and pervasive. Rules, regulations and authorities must be put in place, and they must maintain a well-balanced and wise position. But also new forms of local governance based on agreements, cooperation and private–public synergy can perform well, and even better than traditional 'government' interventions. The various categories of territorial capital are set out in Figure 7.2 and then described.

a) Public goods and resources

Traditional public goods are social overhead capital and infrastructure, natural and cultural public-owned resources, and environmental resources. They are at the basis of the general attractiveness of the local territory, and they represent externalities which enhance the profitability of local activities. Two factors limiting the full exploitation of these resources may be pointed out: unsustainable exploitation and increasing land rents which appropriate a large share of potential profits. Counterbalancing elements and policies in these cases may be: enforced regulations on resource or land use and 'polluter pays' taxation in the case of environmental or landscape damage; taxation with earmarking for resource maintenance and accessibility in the case of land rents.

b) Intermediate, mixed-rivalry tangible goods

Intermediate mixed-rivalry goods are, namely: proprietary networks in transport, communication and energy; public goods subject to congestion effects; collective goods made

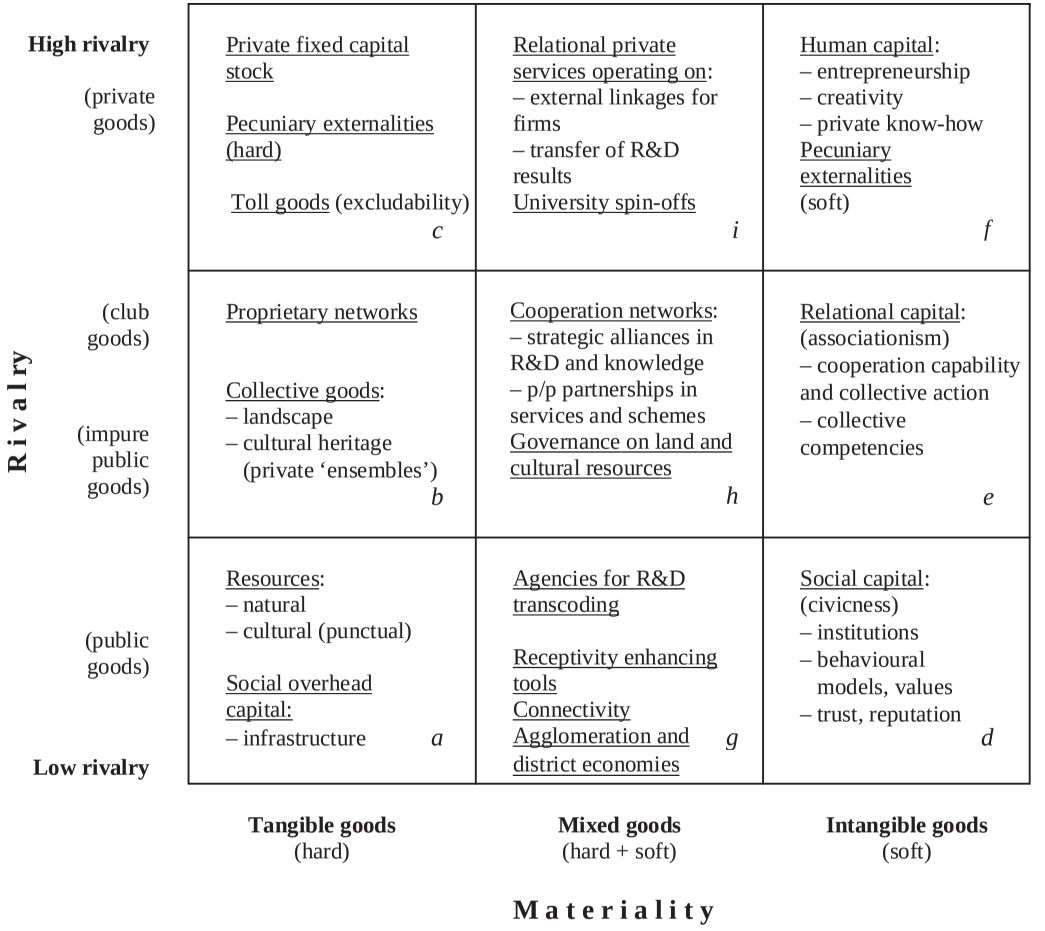


Figure 7.2 A theoretical taxonomy of the components of territorial capital

up of a mix of public and private-owned goods like the urban and rural landscape, or complementary assets defining a cultural heritage system. The first category is generally subject to a control authority guaranteeing fair access, the absence of monopoly pricing, and sufficient maintenance and innovation of the network or good. The last two categories deserve closer inspection: they mainly comprise public or collective goods subject to congestion or free-rider effects that require a mix of control and incentive measures in order to maintain the potential beneficial externalities that they may supply.

In these cases, careful, far-sighted and sustainable private use (or complementary use) of the resource is necessary, and game theory does not allow us to exclude short-term, opportunistic behaviour by some users (or property owners) (Greffe, 2004). In the case of the strict complementarity of single private goods (for example a historic city centre represented by multiple properties and a mix of private and public goods), the long-term advantage of cooperative behaviour is clear; but awareness of this fact depends on the cultural and economic homogeneity of the property owners. Here, a strong sense of belonging and territorial loyalty coupled with a far-sighted business perspective and the social stigmatisation of opportunistic behaviour – the 'milieu' effect – may result in favourable collective action, easy public–private agreements and fruitful local synergies (Camagni et al., 2004). In this case, the milieu itself may be the true territorial capital allowing long-term efficiency in the economic exploitation of local resources (see typology *e)* in the taxonomy).[4]

c) Private fixed capital and toll goods

Private fixed capital stock is, of course, a traditional component of territorial capital. In the short term it may be considered a territorial endowment which enables advantage to be taken of expansions in world trade demand; in the longer run it may be volatile and mobile, although it may be anchored to the local realm by softer but characteristically local and less mobile factors like skills, entrepreneurship and knowledge. In the same class one may also place pecuniary externalities, of a hard nature, encompassing high-quality capital goods or intermediate goods produced in the local context and sold on the market.

A third category, already mentioned, comprises public but tolled goods, in particular when the tolls fully cover construction and maintenance costs.

d) Social capital

To be found (on the side of intangible goods) in our taxonomy, still of a public or collective nature, is social capital. The concept (Coleman, 1990; Putnam, 1993; Grootaert and van Bastelaer, 2001) may now be considered sufficiently established, but its economic nature and its components still do not find sufficient consensus among scholars. Social capital can be defined as the set of norms and values which govern interactions between people, the institutions into which they are incorporated, the relational networks established among various social actors, and the overall cohesion of society. In a word, social capital is the 'glue' that holds societies together.

For economists it includes the capital represented by the rules, habits and relationships which facilitate exchange and innovation, with the consequence that it affects economic development. It is in fact almost unanimously accepted that if a market is to function properly, it needs shared norms as well as institutions and modes of behaviour which reduce the cost of transactions, which ensure that contracts are observed and implemented, and which can rapidly resolve disputes.[5]

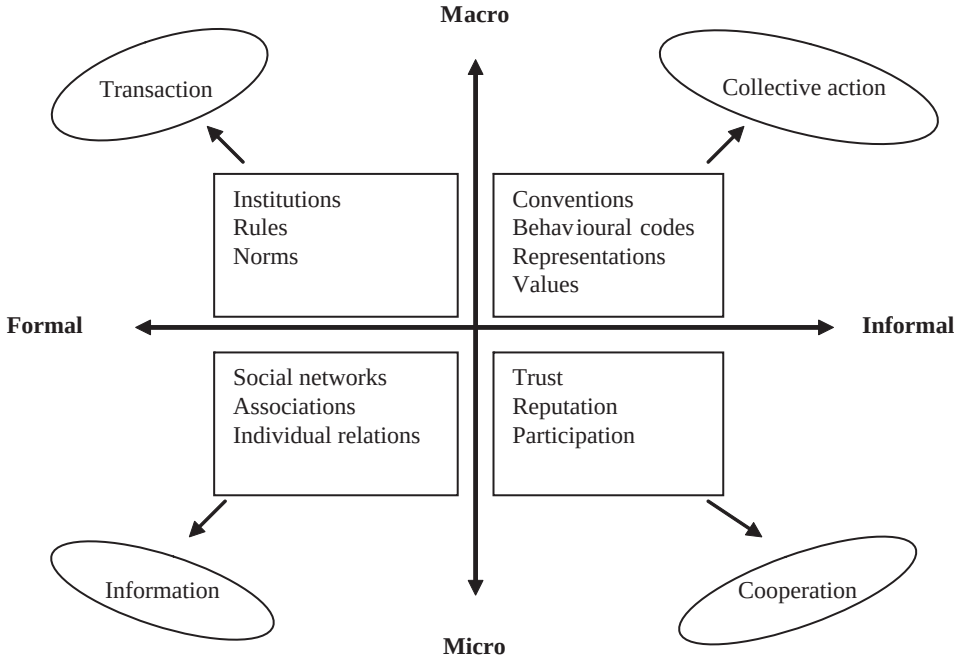
However the concept of social capital has difficulties and ambiguities of an analytical and linguistic nature which still obstruct its full acceptance. The term 'capital' denotes that it is an asset, or stock, accumulated over time which generates a flow of benefits, not just a set of values and social organizations. As a consequence, it should be possible to show that it is built up through a process involving costs or investments, at least in terms of individual and organisational time and effort.[6] On the other hand, social capital is created and accumulated through slow historical processes, and its original function is not directly linked to economic goals, namely an increase in economic efficiency. Therefore, it may be seen as 'a by product of a pre-existent fabric of social relationships, oriented to other goals' (Bagnasco, 2002, p. 274). Rather than being a measurable input to add to other factors of production, it can be considered a public good that produces externalities for the entire economic system, increasing the efficiency of the other factors. From this perspective it would be more appropriate to equate social capital with another well-known economic variable: the level of technological knowledge which, in a production function, moves 'total productivity' of production factors upwards (Camagni, 2004).

In order to avoid an excessively broad definition of social capital, and its use as a 'catch-all' term, it seems helpful to set out a classification of the different components of social capital according to two dimensions, or relevant dichotomies: the micro–macro dichotomy, which distinguishes elements directly involving single individuals from those of the system, and the dichotomy between the formal and the informal dimension, distinguishing elements expressed through observable objects (roles, networks, norms or social structures) from more abstract elements such as values, representations, attitudes and codes of behaviour (Figure 7.3).

The macro dimension comprises institutions and rules in the sense of Williamson and North: 'the rules of the game in a society or, more formally, the humanly devised constraints that shape human interaction' (North, 1990, p. 3). They may be formally expressed and objectively defined, or they may be informal, and here the reference is to conventions, codes of behaviour, values and representations. The micro dimension comprises – among the formal elements – social networks and associations, the ability to focus and organise within organised structures (even loose structures), a large range of interactions among social actors, as well as individual relationships, seen as the set of relations and contacts an individual possesses and which may be invested in economic and social activity. Among the informal elements, however, are trust and reputation and all the non-structured forms of individual participation in public or collective decisions.

There are many channels through which the different elements of social capital may affect local development. At the risk of oversimplifying the theoretical framework, we may state that each case has a more direct role in a specific direction, as indicated in the ovals of Figure 7.3.

Institutions, rules and norms, in fact, fairly explicitly aim to reduce transaction costs, or the use costs of the market. They provide guarantees for contracts and obligations, efficiently manage problems of company law and governance, monitor for conflicts of interest and monopoly practice; in short, they create a favourable business climate which benefits local firms and enhances attractiveness for external firms. Social networks and associations aim to reduce the costs (and increase the availability) of information, particularly for current and potential commercial partners. They widen the potential market, make it easier to identify and sanction opportunistic behaviour, and accelerate the



Source: Camagni (2004).

Figure 7.3 Dimensions, forms and roles of social capital

transmission of information on good practices, thereby facilitating their imitation and diffusion. Conventions and common values allow collective action among private parties to be undertaken more easily, that is, the *ex ante* coordination of individual decisions in order to achieve the advantages of economies of scale, purpose and complementarity. In many cases it is only if decisions are taken concurrently that costs can be reduced and complex projects made profitable and viable. Trust and reputation facilitate exchanges and repeated contracts, cooperation (covenants, strategic alliances, contracts – even incomplete – between customers and suppliers) or partnerships between public and private parties.

In all cases, the importance of social capital for economic activity is entirely evident.

The macro elements charted in the upper part of the figure refer to civicness and are appropriately indicated in the bottom-right cell of Figure 7.2; on the other hand, the micro elements of social capital refer to associationism and social networks and are more appropriately shown in cell *e* of Figure 7.2, devoted to what we call 'relational capital'.

e) Relational capital

Social capital may be given either a 'systemic' or a 'relational' interpretation according to the generality of the approach, the emphasis on a 'general-purpose' versus a 'selective' interpretation of its economic role, and the attention paid to economic potential versus actual economic outcome. While it may be argued that a social capital exists wherever a society exists, 'relational' capital may be interpreted as the set of bilateral and

multilateral linkages that local actors have developed, both inside and outside the local territory, facilitated by an atmosphere of easy interaction, trust, shared behavioural models and values. In this sense, relational capital is equated with the concept of the local milieu, meaning a set of proximity relations which bring together and integrate a local production system, a system of actors and representations and an industrial culture, and which generates a localised dynamic process of collective learning (Camagni, 1991). Geographic proximity is associated with socio-cultural proximity – the presence of shared models of behaviour, mutual trust, common language and representations, common moral and cognitive codes.

The role of the local milieu, and consequently of relational capital, in terms of economic theory is linked to three types of cognitive outcome which support and complete the normal mechanisms of information circulation and coordination of agents performed through the market: namely, reduction of uncertainty in decisional and innovative processes through socialised processes of information transcoding, imitation and control among potential competitors; *ex ante* coordination among economic actors facilitating collective action; and collective learning, a process occurring within the local labour market and which enhances competencies, knowledge and skills.[7]

In public–private terms, relational capital and milieu effects belong to an intermediate class where 'collective' rather than public efforts and investments give rise to beneficial effects that can be exploited only by selectively chosen partners located in particular territories with specific identities, and sharing similar interests and values. The concept of club goods seems best suited to interpreting this condition.

f) Human capital

The presence of human capital is today constantly cited as a fundamental capital asset available to territories so that they can compete in the international arena by both strengthening local activities and attracting foreign ones. Endogenous growth theories long since developed the concept into formalised growth models (Lucas, 1988; Romer, 1990), thereby starting a significant and fruitful process of convergence between stylised approaches and qualitative, bottom-up development theories (Capello, 2007). In parallel, the concept of territorial capital, once it has been duly developed and analytically structured, could become the attractor and the interlocking element between the two theoretical trajectories – endogenous growth and endogenous development theories.

Besides human capital, this class also comprises the pecuniary externalities supplied by the territorial context in terms of advanced private services in the fields of finance, technological and marketing consultancy, customized software packages, and so on.

g) Agglomeration economies, connectivity and receptivity

Again belonging to the class of public or collective goods of a mixed – hard and soft – nature are those elements of territorial capital that concern:

- Agglomeration economies, or – in different territorial contexts characterised by specialisation in some sectors, technologies or *filières* – district economies. Cities and industrial districts, viewed as archetypes of the territorial organisation of production and social interaction, exhibit clear similarities in theoretical terms in spite of their geographical and economic differences (proximity and high density of

activities, concentration of social overhead capital, density of interaction, high cohesion and sense of belonging) (Camagni, 2004). These similarities give rise to economic advantages like the reduction in transaction costs, cross-externalities, division of labour and scale economies that constitute a large part of territorial capital.

- Connectivity, by which is meant the condition in which pure physical accessibility is utilised in a targeted and purposeful way by the single actors in order to collect information, organise transactions and exchange messages in an effective way.
- Receptivity, or the ability to extract the highest benefit from access to places, services or information.
- Transcoding devices, operating in the field of knowledge accumulation and diffusion, mainly in the form of public agencies facilitating interaction among research facilities, universities and firms and whose mission is to create a common language and shared understanding among the above-mentioned bodies.

h) Cooperation networks

This category of territorial capital lies at the centre of the 'innovative cross'. It integrates tangible and intangible assets and yields goods and services traditionally supplied through public–private or private–private cooperation networks. Strategic alliances for R&D and knowledge creation supported by (or partially supporting) public agencies for the dissemination and diffusion of knowledge, operating on the open market with some public support, are the key tools for a fair and fast implementation of the knowledge society.

But the advantages of a public–private partnership strategy do not reside only in management of the knowledge *filière*. The strategy also allows crucial potential results to be achieved by urban schemes for the development of large urban functions and services (where *ex ante* coordination among partners enhances private profitability and public efficiency in the investment phase).

A third area in which this class of territorial capital is manifest consists of new forms of governance in spatial planning and land use, a field characterised by both market failures and government failures, but also by huge risks of contradictory strategies and undesirable outcomes if individual, piecemeal, non-cooperative private decisions are not controlled (OECD, 2001).

In all the cases mentioned above, the term 'capital' can be used on sound economic bases: the construction of relational networks and cooperation agreements involves real and costly investments which are usually overlooked owing to their nature as implicit or sunk costs (management time, organisational costs, risk of failure or of opportunistic behaviour by potential partners) (Camagni, 1993).

i) Relational private services

Of course, in many cases certain crucial services of a relational nature may be supplied entirely by the market: for example, when firms search for external partners and suppliers (through financial institutions or specialised consultancy agencies), or in the cases of technological transfer, partnership and diffusion. University spin-offs also belong to the class of potential territorial assets to be supported by internal rules and public incentives – financial or 'real'.

7.4 Conclusions

It appears from the foregoing discussion that territorial capital is a new and fruitful concept which enables direct consideration to be made of a wide variety of territorial assets, both tangible and intangible, and of a private, public or mixed nature.

These assets may be physically produced (public and private goods), supplied by history or God (cultural and natural resources, both implying maintenance and control costs), intentionally produced despite their immaterial nature (coordination or governance networks) or unintentionally produced by social interaction undertaken for goals wider than direct production. In all cases, a repeated use in successive production cycles of these assets is implied, and the usual accumulation–depreciation processes take place – as in the case of physical capital assets. In most cases, the accumulation process is costly, except when socialised processes taking place within the territorial context are responsible for the cumulative creation and value of an immaterial asset.[8]

The economic role of territorial capital is to enhance the efficiency and productivity of local activities. A stylised, potential treatment of the single elements of territorial capital should address its efforts towards finding a way to measure each of them quantitatively. The impossibility of direct measurement implies equating the effects of territorial capital with 'technological progress' in a production function – but this would only be a measure of our 'ignorance'.

In this chapter, a preliminary taxonomy of the various components of territorial capital is proposed. Based on the two dimensions of rivalry and materiality, this taxonomy has gone beyond the traditional 'square' encompassing pure private and pure public goods, human capital and social capital. An intermediate class of club-goods or impure public goods has emerged which implies, or requires, strong relationality and seems of great relevance to the governance of local development processes. On the one hand are proprietary networks – of a hard nature when they are physical, or a soft nature when they concern cooperation agreements and the supply of common services; on the other, there are public goods subject to congestion or to opportunistic, free-rider or endangering behaviour. In both cases, new forms of governance, participatory and inclusive, should be developed in order to accomplish the maximum benefit for the members of the 'club' – the local community. The presence of social or relational capital in the form of trust and cooperative attitudes is highly beneficial in this respect.

Generally, tangible assets are subject to traditional supply processes, while intangible assets operate in the sphere of 'potentials'. The 'mixed' category, which merges the two components together, translates abstract potentials into actual assets by defining shared action strategies, complex relational services and concrete cooperation agreements between private and public partners.

The 'mixed' category of 'hard+soft goods' has the further advantage of highlighting the relevance of such complex territorial organisations as cities or 'districts'. These are sorts of collective goods built through the spontaneous, unorganised action of a multitude of local actors, private and public, and which thus generate wide externalities for the entire community. Once again, wise control policies should be implemented in order to avert the implicit risk of rent-seeking behaviour: the localised nature of these public goods automatically generates increases in land rents which, on the one hand, may be beneficial in that they trigger a continuous upward selection process in the quality of local activities and a 'filtering-down' process of lower-order functions along the urban

hierarchy, but on the other hand subtract potential profits from productive (social classes and) uses.

All the above considerations have significant implications for new spatial development policies (OECD, 2001; Camagni, 2001) which introduce governance styles addressed to cooperation and relationality. A telling example of the style required is provided by the new strategies necessary to cope with the issue of the knowledge society: instead of (or besides) injecting public money directly into the system of firms, universities and research centres, which by and large are self-referential systems with their own specific goals, public policy should support 'relational' actions, such as common schemes and production projects built through cooperation among the above-mentioned actors operating on the local or regional scale; or it should support 'transcoding' services linking scientific output and business needs and ideas, such as transfer of R&D, development of a science-based entrepreneurship or university spin-offs. More generally, the approach suggests a new role for local or regional policy-makers as the 'facilitators' of linkages and cooperation among actors, at both the regional and the interregional and international scale.

Notes

1. We shall find that, on an *ex post* base, the national aggregate growth rate and the weighted sum of regional growth rates are equal.
2. Every country always has a 'comparative advantage' in some production sectors, even if it may be less efficient in absolute terms in all production with respect to competitor countries: its advantage resides in production in which it is 'comparatively' less inefficient, and it is exactly in such production that it will specialise within the international division of labour, to the mutual benefit of all countries. The Ricardian principle of comparative advantage was judged by Paul Samuelson as the only statement of economic theory that was at the same time true and not trivial. As argued here, it refers to countries, not to regions or territories (see also Camagni, 2001).
3. GREMI: Groupe de Recherce Européen sur les Milieux Innovateurs, headquartered in Paris at Université de Paris 1 – Panthéon Sorbonne and active since the mid-1980s.
4. Does all this mean that the local milieu is per se an ethical and environment-friendly subject or intermediate institution? The answer is no, of course: a lobbying and short-term strategy by local, situation-rent seeking actors is not excluded, if not probable, and a mix of regulations and incentives implemented by public bodies seems necessary. In the case of external challenges and threats to local business, the presence of a milieu guarantees a faster and more effective reaction capability (see Camagni and Villa Veronelli, 2004, describing the case of an apple-producing community in the Trento Valley, Italy, challenged by the anti-pesticide health regulations imposed in their major German market).
5. If we add further factors – reciprocal trust, a sense of belonging to a community that shares values and behaviours, and participation in public decisions – then a climate is created which encourages responsibility, cooperation and synergy. Such a climate enhances productivity, stimulates creativity and ensures more the effective provision of public goods.
6. This is the rationale of research programmes which attempt to measure social capital by using suitable proxies (Putnam, 1993; Arrighetti et al., 2001) so as to include it in an ideal production function along with human capital and physical capital.
7. Also to be mentioned here is the function of promoting informal guarantees for the honouring of incomplete contracts, which the milieu can perform because of its networks of interpersonal relations. Models inspired by game theory have been used to show that, when there are interpersonal networks and effective mechanisms for punishment, social exclusion and reprisal, implying a reduction in the costs of monitoring and enforcement of contracts, it is possible not only to attain stable (cooperative) Nash equilibria which are not possible when costs are high, but also to achieve overall benefits for the partners which exceed the allocative costs of local contractual policies (or 'parochialism') (Bowles and Gintis, 2000).
8. This feature is also present in the case of physical, costly capital assets – for example the effects of increasing agglomeration externalities on the value of real estate assets.

References

Arrighetti, A., A. Lasagni and G. Seravalli (2001), 'Capitale sociale, associazionismo economico e istituzioni: indicatori statistici di sintesi', Working Papers no. 4, Dipartimento di Economia, Università di Parma.

Aschauer, D. (1989), 'Is public expenditure productive?', *Journal of Monetary Economics*, **23**, 177–200.

Bagnasco, A. (2002), 'Il capitale sociale nel capitalismo che cambia', *Stato e Mercato*, **2** (65), 271–303.

Becattini, G. (1990), 'The Marshallian industrial district as a socio-economic notion', in F. Pyke, G. Becattini and W. Sengenberger (eds), *Industrial Districts and Inter-firm Cooperation in Italy*, Geneva: ILO.

Biehl, D. (1986), 'The contribution of infrastructure to regional development', document, Commission of the European Communities, Bruxelles.

Bowles, S. and H. Gintis (2000), 'Optimal parochialism: the dynamics of trust and exclusion in networks', Department of Economics, University of Massachusetts, February, mimeo.

Camagni, R. (1991), 'Local *milieu*, uncertainty and innovation networks: towards a new dynamic theory of economic space', in R. Camagni (ed.), *Innovation Networks: Spatial Perspectives*, London: Belhaven-Pinter, pp. 121–44.

Camagni, R. (1993), 'Interfirm industrial networks: the costs and benefits of cooperative behaviour', *Journal of Industry Studies*, **1**, 1–15.

Camagni, R. (1999), 'The city as a *milieu*: applying GREMI's approach to urban evolution', *Revue d'Economie Régionale et Urbaine*, **3**, 591–606.

Camagni, R. (2001), 'Policies for spatial development', in OECD, *OECD Territorial Outlook*, Paris: OECD, pp. 147–69.

Camagni, R. (2002), 'On the concept of territorial competitiveness: sound or misleading?' *Urban Studies*, **13**, 2395–2412.

Camagni, R. (2004), 'Uncertainty, social capital and community governance: the city as a milieu', in R. Capello and P. Nijkamp (eds), *Urban Dynamics and Growth: Advances in Urban Economics*, Amsterdam: Elsevier, pp. 121–52.

Camagni, R. and R. Capello (2002), 'Milieux innovateurs and collective learning: from concepts to measurement', in Z. Acs, H. de Groot and P. Nijkamp (eds), *The Emergence of the Knowledge Economy: A Regional Perspective*, Berlin: Springer-Verlag, pp. 15–45.

Camagni, R. and D. Maillat (eds) (2006), *Milieux innovateurs: théorie et politiques*, Paris: Economica.

Camagni, R. and D. Villa Veronelli (2004), 'Natural resources, know-how and territorial innovation: the apple production system in Val di Non, Trentino', in R. Camagni, D. Maillat and A. Matteaccioli (eds), *Ressources naturelles et culturelles, milieux et développement local*, Neuchâtel: Editions EDES, pp. 235–60.

Camagni, R., D. Maillat and A. Matteaccioli (eds) (2004), *Ressources naturelles et culturelles, milieux et développement local*, Neuchâtel: Editions EDES.

Capello, R. (1994), *Spatial Economic Analysis of Telecommunication Network Externalities*, Aldershot: Avebury.

Capello, R. (2001), 'Urban innovation and collective learning: theory and evidence from five metropolitan cities in Europe', in M.M. Fischer and J. Froehlich (eds), *Knowledge, Complexity and Innovation Systems*, Berlin, Heidelberg and New York: Springer, pp. 181–208.

Capello, R. (2007), *Regional Economics*, London: Routledge.

Coleman, J.S. (1990), *Foundations of Social Theory*, Cambridge, MA: Harvard University Press.

Dosi, G. (1982), 'Technological paradigms and technological trajectories', *Research Policy*, **3**, 147–62.

European Commission (2005), 'Territorial state and perspectives of the European Union', Scoping document and summary of political messages, May.

Foray, D. (2000), *L'Economie de la Connaissance*, Paris: La Découverte.

Gilly, J.P. and A. Torre (eds) (2000), *Dynamiques de Proximité*, Paris: L'Harmattan.

Greffe, X. (2004), 'Le patrimoine dans la ville', in R. Camagni, D. Maillat and A. Matteaccioli (eds), *Ressources naturelles et culturelles, milieux et développement local*, Neuchâtel: Editions EDES, pp. 19–44.

Grootaert, C. and T. van Bastelaer (2001), 'Understanding and measuring social capital: a synthesis of findings and recommendations from the social capital initiative', World Bank, Social Capital Initiative Working Paper no. 24, April, Washington, DC.

Lucas, R. (1988), 'On the mechanics of economic development', *Journal of Monetary Economics*, **22**, 3–42.

Malmgren, H.B. (1961), 'Information expectation and the theory of the firm', *Quarterly Journal of Economics*, **75**, 399–421.

Nelson, R. and S. Winter (1982), *An Evolutionary Theory of Economic Change*, Cambridge, MA: Harvard University Press.

North, D. (1990), *Institutions, Institutional Change and Economic Performance*, Cambridge: Cambridge University Press.

OECD (2001), *OECD Territorial Outlook*, Paris: OECD.

Putnam, R.D. (1993), *Making Democracy Work*, Princeton, NJ: Princeton University Press.

Romer, P. (1990), 'Endogenous technological change', *Journal of Political Economy*, **98**, S71–S102.

Simon, H. (1972), 'From substantive to procedural rationality', in C.B. McGuire and R. Radner (eds), *Decision and Organization*, Amsterdam: North-Holland.
Storper, M. (1995), 'The resurgence of regional economies ten years later: the region of untraded interdependencies', *European Urban and Regional Studies*, **2**, 191–221.
Storper, M. (2003), 'Le economie locali come beni relazionali', in G. Garofoli (ed.), *Impresa e territorio*, Bologna: Il Mulino, pp. 169–207.
Williamson, O. (2002), 'The lens of contract: private ordering', *American Economic Review, Papers and Proceedings*, **92** (2), 438–53.

8 Human capital and regional development

Alessandra Faggian and Philip McCann

8.1 Introduction

Over the last two decades, the analysis of human capital has taken on a progressively more central role in discussions regarding the growth and success of nations and regions. This is primarily because advanced societies have increasingly evolved towards what has been called a 'knowledge-based economy' (OECD, 2006), whereby tertiary-level human capital is seen to be a crucial feature of economic growth. However, the links between human capital and national economic development may not necessarily be the same as those between human capital and regional economic development. The reason for this is that there are two quite distinct sets of human capital impacts on regions, the first of which mirrors the national impact, while the second differs markedly from the national impact. Firstly, as with national economies, the human capital in a region has an impact on the aggregate productivity in the economy, via the externalities associated with it. Secondly, and rather differently to national economies, human capital in a region can also result in a major spatial reallocation of factors. These two impacts do not always operate in the same direction, because the mechanism by which externalities spill over into a local region and the mechanisms which determine labour mobility are not necessarily congruent. In circumstances where these two impacts coincide, regions will flourish, whereas in situations where they do not coincide, regions will struggle.

As we will see in this chapter, the impacts of human capital on regional development depend on the link between these two mechanisms. Yet, since the 1950s, our understanding of these issues has changed. In advanced economies, the early analysis of worker know-how and skills, collectively referred to as 'human capital', emerged in the 1950s and 1960s primarily with a microeconomic focus on individual behaviour. For non-spatial analyses the focus was on returns to educational investments, whereas for explicitly spatial analyses, the focus was on human capital–migration interactions. This remained the case until the late 1980s, when the advent of 'new growth theory' changed our understanding of the links between human capital and aggregate economic development. This reformulation of neoclassical growth framework by the new growth theories also encouraged a reconsideration of the links between human capital and regional development, and the major issue to come out of this was reconsideration of the links between labour migration and local economic growth. However, for regional development issues, our analyses have recently been further complicated by the fact that the notion of human capital has been extended and redefined in a variety of ways, many of which are much broader than the original usage. There is therefore currently some ambiguity regarding how this term is most appropriately used and how the links between human capital and regional development are measured empirically.

In order to understand these issues we begin in section 8.2 by analysing the non-spatial impacts of human capital on economic development, and we then explicitly examine in section 8.3 the 'spatial' impacts of human capital by introducing the role of interregional

migration. Section 8.4 clarifies the nature of the various theoretical links between human capital regional growth models and interregional migration of human capital. Section 8.5 discusses the recent empirical evidence for these links, with a particular focus on the recent resurgence of cities in both attracting and fostering human capital. In section 8.6 we extend the argument to additional and related notions of human capital, and discuss the relationship between human capital and creativity. Section 8.7 provides some brief conclusions.

8.2 Human capital and economic growth

Many of the concepts which emerged as being central to the human capital literature were initially anticipated by Friedman and Kuznets (1954), although the first formal analysis of the returns to schooling was conducted by Mincer (1958). From the 1960s onwards, however, the centrality of the notion of human capital to economic theory was in large part due to Gary Becker, whose 1964 book *Human Capital* provided the first unifying framework on the subject. In this framework, the two most identifiable ways of investing in human capital are education and on-the-job training (Becker, 1964), and the wage premia earned by these two components of human capital are estimated using a Mincer (1974) earnings function. Due to data limitations, the number of years of education is normally used as the best proxy for human capital, with age adjustments and years of employment being incorporated into the earnings equations. However, while the traditional focus of the human capital literature tended to be on secondary education in advanced economies and on both primary and secondary education in developing economies, more recently the focus of education research and policy in the industrialised economies has moved towards tertiary education in response to the technological transformations that have been taking place. This is because it has become increasingly clear that high levels of secondary-level education are insufficient to compete in the new global economy, and that tertiary education skills are instead required. Between the mid-1980s and the mid-1990s increased enrolment in tertiary education was accompanied by a substantial expansion of the knowledge sector. According to OECD (2001), between 1986 and 1994 the value-added in knowledge-based industries in Organisation for Economic Co-operation and Development (OECD) countries grew an average of 3 per cent versus 2.3 per cent for the business sector overall. Higher education is therefore seen nowadays as playing an increasingly crucial role in a country's economic well-being and development, because only higher-level education and skills are perceived as being sufficient to allow countries to compete in the globalised knowledge sectors. The result has been that since the 1970s, the general trend across the developed world has been to increase participation in higher education as a way of investing in human capital. The expansion has also affected developing countries, although in these countries the expansion rate has been slower, increasing the gap between the developing countries and the industrialised countries.

In terms of regional development issues, the critical role played by tertiary education in advanced economies, rather than primary or secondary education, has led to a focus on the interactions between higher education institutions (HEIs) and their local regions. In this context, the dominant approach has been the use of 'regional multipliers'[1] (with the only exception of Felsenstein,[2] 1997) as a way of measuring the local and regional income–expenditure–employment effects associated with the HEI. However, while it is

true that universities are beneficial to local economies because of 'multiplier effects', regional multipliers are only part of a very much larger story. As initially pointed out by Sjaastad (1962), the reason for this is that migration is also a means by which individuals can acquire human capital or reap the rewards to human capital, and this observation would appear to be particularly pertinent to university graduates. Therefore, the long-term growth and allocative impacts of tertiary-educated human capital on regional development are probably far more important than the short-term multiplier effects of HEIs.

In order to understand the links between human capital and regional development, we must therefore first consider the role of tertiary education within the overall context of aggregate growth models. Secondly, we must seek to identify the ways in which these non-spatial models can be adapted and interpreted within a regional and spatial context, by focusing on the issue of human capital migration. In the rest of this section we focus primarily on the non-spatial analysis and we subsequently develop our explicitly spatial interpretation in the following sections.

In terms of non-spatial analyses, according to the traditional neoclassical growth model (Solow, 1956; Swan, 1956), growth can be explained by the interaction between traditional inputs such as capital and labour, and an exogenous unexplained generic input often referred to as 'total factor productivity' or 'the residual'. It was assumed that this unexplained generic input is maximised in an environment of efficient markets, such that growth as a whole is maximised by the efficient allocation of resources. Over time, however, it became increasingly apparent that this unexplained portion of the growth rate was actually in many cases the largest and most important part of economic growth. Neoclassical economists assumed the residual to be a product of technological innovation, but innovation itself remained something undefined within the neoclassical framework.

It was Romer (1986, 1990, 1994) who first pointed out that this technological progress can actually be 'endogenised' in the production function so that the total output is in fact the product of three (and not two as in the pure neoclassical framework) inputs. Mathematically (and assuming the standard Cobb–Douglas aggregate production function typical of the neoclassical models) this was expressed as:

$$Y = K^{\alpha} L^{1-\alpha} \mathbf{K}^{\beta} \tag{8.1}$$

where K and L are the inputs used by the single firm, which have diminishing returns to scale. $\mathbf{K}$ is the aggregate stock of capital, which is equal to nK the number of firms multiplied by the capital belonging to each of them. If β proves to be significantly greater than zero, then the Cobb–Douglas production function has increasing returns to scales. The growth rate of the system grows in time and the conclusion is the opposite of the neoclassical 'catching-up' process. Romer's new growth theory highlighted the possibility of introducing an endogenous source of growth in a simple neoclassical model without abandoning the framework of perfect competition. As a consequence, a plethora of studies appeared, trying to redefine the 'third input' in the production function, as this represented the real 'engine of growth'. Among these, Lucas (1988) identified 'human capital' as a possible explanation for endogenous growth. He proposed the following version of the aggregate production function:

$$Y = AK^{\beta}(uhL)^{1-\beta} h_a^{\gamma} \tag{8.2}$$

where K is physical capital, L is the number of workers, h is human capital, and $(1-u)$ is the number of hours per day each worker devotes to learning activity. In addition, the development of the complete model also employs the following parameters: β, γ, σ, ρ, and ϕ, whereby ρ is the rate of time preference and σ is the inter-temporal elasticity for the substitution of consumption.

In equation (8.2) human capital appears twice. The first time it represents the effect on the productivity of each worker, and the second time it represents the positive externality on the productivity of the whole economy. Lucas (1988) called the first effect 'internal' and the second 'external'.

Lucas's model also reaches the conclusion that the market efficient rate of human capital growth is lower than the optimal rate for society, because of externalities due to the average level of human capital h_a. Calling v the rate of human capital growth in the first case and v^* in the second, it can be demonstrated that:

$$v = \frac{(1-\beta)(\phi-\rho)}{\sigma(1-\beta+\gamma)-\gamma} \tag{8.3}$$

and:

$$v^* = \frac{(1-\beta)(\phi-\rho)+\phi\gamma}{\sigma(1-\beta+\gamma)} \tag{8.4}$$

In the absence of externalities, γ is equal to zero and the rates are the same. On the other hand, if positive externalities exist, then $\gamma > 0$ and $v^* > v$. The growth rate of the economy, g, is given by:

$$g = \left(1 + \frac{\gamma}{1-\beta}\right) v \tag{8.5}$$

As such, the Lucas (1988) model demonstrates that a higher level of human capital allows the economy to grow faster and the inputs to be better paid, as long as positive externalities are associated with the average level of human capital.

The potential of endogenous growth models à la Lucas to study regional development is evident. Being able to identify the 'engine of growth' in a region is as important as, or even more so, than in a national context, to foster development. However, these models were conceived to study a closed economy as a first approximation, but regions are open economies and the interactions among regions are likely to be even more crucial than those among countries. The smaller the context of application, the greater is the importance of flows of goods, people and information towards and from other areas. So far, not many convincing applications of new growth theory at a regional level have been published and the general feeling is that there is still a lot of work to do in this field. One exception is Nijkamp and Poot (1998) who point out that 'at the regional level, there are spatial interactions in term of trade, capital flows, migration, diffusion of technological innovation and information exchanges. Thus, the closed economy models can provide at best a very limited understanding of regional growth'. In their contribution, they analyse the impacts of labour and capital migration, and diffusion of technology and trade, on regional economic development by adding new equations to the neoclassical model. One

of their results is that immigration tends to lower growth rate, unless immigrants are highly skilled workers. This is a crucial point and one that so far has very often been overlooked in the literature.

Many other contributions in very recent years have tried to apply endogenous growth models to the regional context, but most of the authors simply use models conceived for a national level to study the regional context, adding some extra variables. The problem is that the nature of the phenomenon under investigation is completely different in the two cases. What is important in a country-level study is not supposed to have the same importance when we study a smaller context, and vice versa. It is not just a matter of different 'scale'. Applying exactly the same methodology with a few changes can therefore be misleading. The spatial dimension is essential. The localisation of a region has a high influence on its economic performance and so do characteristics of the region, such as its industrial mix and infrastructures. More fundamentally, it is also vital to study in detail the kind of interactions the region has with other areas. In this respect, we argue that the role of interregional migration, especially of highly educated people, should be central in the study of regional development. Section 8.3 will therefore be focused on this phenomenon.

8.3 The migration behaviour of human capital

While we know that economic growth is positively related to human capital, the application of the principles of endogenous growth macroeconomic models à la Lucas (1988) to regions is not at all straightforward. There are at least two reasons for this. The first, which has been extensively treated in the regional literature and it is at the base of the regional economics discipline itself, is that regions are inherently very different from nations. Not having proper boundaries, regions have a much higher degree of 'openness' and factors can flow relatively easily between them. In particular, labour is potentially very mobile between regions within the same country. The second reason, strictly connected to the first one, is that labour migration behaviour is itself dependent on the level of individual human capital. One of the most uncontroversial results in migration studies is, in fact, that embodied human capital (mainly in the form of education) significantly increases the probability of migrating (Bartel, 1979; Bogue, 1985; DaVanzo, 1976, 1978, 1983; Faggian and McCann, 2006; Faggian et al., 2006; Faggian et al., 2007a, 2007b; Plane and Rogerson, 1994; Polachek and Horvath, 1977; Polachek and Siebert, 1993; Schwartz, 1973, 1976; Stark and Taylor, 1991; Topel, 1986; Rebhun, 2003). The migration of individuals with high 'embodied human capital' introduces a complication in the human capital–regional development relationship because the increase in human capital due to education can easily leak out of an area even when produced there, and therefore not generate the forecasted 'multiplier effects'.

Several explanations for the greater mobility of the educated have been suggested, but they can all be linked either to the so-called 'human capital migration theory' or to the 'job search theory'.

The 'human capital migration theory' dates back to the seminal works of Sjaastad (1962) which treats both migration and acquiring education as investment decisions. As such, migration, like any other investment, has associated costs and returns. A person will decide to migrate only when the net present value of a migration investment is positive. Following Hart (1975) this can be expressed in more formal terms as follows.

Let us suppose that a potential migrant wants to move from region i to region j. The migrant will migrate only if the expected value of utility derived from the net present value of his expected returns (R_i) in the origin region i (origin) is less than that which can be earned in the destination region j minus the costs associated with relocation (C_{ij}):

$$E\{U[R_i](0)\} < E\{U[R_j](0)\} - E\{C_{ij}(0)\} \tag{8.6}$$

where the zero in parenthesis simply means that earnings and cost are evaluated at the present time ($t = 0$). If we assume that the subjective discount rate is r and that migration happens at time $t = T$, equation (8.7) becomes:

$$\int_0^T e^{-rt} U[R_i(t)]dt < \int_0^T e^{-rt} U[R_j(t)]dt - \int_0^T e^{-rt}[C_{ij}(t)]dt \tag{8.7}$$

Costs and returns can be classified into private and social. The former are, in turn, divided into monetary and non-monetary. The private monetary costs of migration, similar to those of education, are out-of-pocket costs. Non-monetary costs include the opportunity costs represented by foregone earnings during the period of travelling, searching for and learning a new job, and the 'psychic costs' (Sjaastad, 1962) due to the fact that people are generally reluctant to leave familiar surroundings, family and friends. As far as returns are concerned, financial returns are in the usual form of higher real wages, while non-money returns reflect the migrant's preference for the new place in terms of amenities such as climate, reduced congestion, pollution and so on.[3]

The basic idea is that highly educated people have lower costs and/or higher returns, so that, overall, it is more likely for them to have a positive future net benefit from migrating. Several reasons have been proposed for this.

First of all, the better educated have lower information costs. Many authors have pointed out the higher effectiveness of better-educated people in accessing information.[4] DaVanzo (1983), for instance, speaks of 'superior ability' of better-educated migrants 'in processing information' and even uses educational attainment as an indicator of the quantity and quality of a person's information about opportunities elsewhere. Levy and Wadicky (1974) underline how the educated have more and better access to information about opportunities in alternative locations, since 'education increases information directly and reduces the cost of obtaining more information'. Furthermore, as Yezer and Thurston (1976) underlined, 'the longer the distance, the more likely unreliable the information'. On one hand education results in a wider search area, but on the other, it has a positive effect on the ability to obtain and analyse information efficiently, thus shortening the search duration.

Another important component of the total cost of moving depends on how strong the links with the place of origin are (often referred to in the literature as psychic costs). Levy and Wadicky (1974) point out the higher 'adaptability' of the educated to new places. They are more 'receptive to change' and therefore less attached to traditional surroundings. DaVanzo and Morrison (1981) notice that, generally, less-educated people seem to have a strong reliance on family and friends. The psychological costs of leaving their origins are therefore higher, sometimes also because of the fewer possibilities they will have to come back to visit relatives and friends in their places of origin, due to their lower incomes and 'budget constraints'. This not only affects their propensity to migrate but

also gives a bias to their choice of destination, should they decide to move. Indeed, when they decide to migrate, they prefer, where possible, to follow relatives or friends who have moved before them. Having a 'network of acquaintances' in the destination is a very valuable means of acquiring knowledge and information about the new place and may help to compensate for their lower efficiency in gathering and processing information.

Another possible reason for the positive association between educational attainment and the propensity to migrate is that there is a sort of 'path-dependency' in the decision to move. Once people decide to leave their place of origin to go and study in a different place, it becomes easier to move again in the future. DaVanzo (1983) is the first to analyse the 'repeat migration'. The decision to migrate is not a 'once and for all' decision. It is not uncommon for people to decide to migrate more than once during their lifetime. As Polachek and Siebert (1993) underline:

> not all job search takes place at a moment in time. In most instances the search continues throughout one's life . . . people often view their job or location as a stepping-stone for further advancement. Searchers can thus be viewed as 'perspicacious peregrinators' because they seek and weigh information on locational and occupational choices in each time period.

Moreover, DaVanzo and Morrison (1981) shows that there are two different kinds of repeat migration: onward moves and return moves. When less-educated people decide to make a repeat move, it is usually a return to the origin to 'correct' a previous movement which ended up being a failure. Conversely, the better educated tend to move on towards new destinations, showing that onward migration offers them an advantageous 'means of reinvesting in human capital'.

Last, but not least, the risk associated with the decision to migrate is lower for educated people. Their chances of being unemployed in the destination are lower because, even if they cannot find the job they want, they can decide to take lower-paying jobs, which are usually available to the uneducated. The reverse does not hold. If we assume, as seems reasonable, that individuals are generally 'risk-adverse', the higher risks associated with less-educated people would reduce their tendency to migrate.

An alternative human capital migration lies in the 'job-search theory'. The standard job-search model was described in the 1970s by Lippman and McCall (1976a, 1976b, 1979) and Pissarides (1976).[5]

The main premise of these models is that searching for a job is a sequential process. In each period a searcher pays a fixed cost (transportation, advertising), denoted as c, and receives a job offer. The job seeker can then decide to continue the search or to accept the previous offer. Each offer is assumed to be a random variable with a cumulative distribution function F, so that the only source of uncertainty in the job-search model is the distribution of job offers. The job seeker's aim is to maximise net benefits, that is, the future stream of income minus search costs, and this is achieved by separating all job offers into two categories: namely acceptable offers and unacceptable offers. The value that divides the two groups is called the reservation wage. Formally, this can be represented as follows. Let us call the reservation wage w and g_w the expected gain from a search with reservation wage equal to w. Let also N_w be the number of offers until an acceptable offer arrives, N_w is a geometric random variable with parameter $p(w) \equiv 1 - F(w)$ and $E(N_w) = 1/p$ then:

$$g_w = -c/p(w) + \int_w^{\infty} y dF(y)/p(w) \tag{8.8}$$

where the term on the right-hand side is simply the expected cost of search plus the conditional expected value of an offer, given that the offer is at least the reservation wage. Setting the first derivative of g_w, with respect to w, equal to zero gives us the following condition for the optimal reservation wage, ξ:

$$c = \int_{\xi}^{\infty} (y - \xi) dF(y) \tag{8.9}$$

The economic interpretation of equation (8.11) is familiar and says that the reservation wage associated with the optimal stopping rule is chosen to equate the marginal cost of search, c, to the expected marginal return from one more job offer.

Other models have subsequently been developed on the basis of this original framework in order to take into account the distribution of wages across regions and their relationship to distances (Rogerson, 1982), changes in the number of available job opportunities associated with business cycle fluctuations, regional unemployment (Jackman and Savouri, 1992), and the issue of whether migration precedes employment or vice versa (Basker, 2003; Fahr and Sunde, 2002; Gordon and Vickerman, 1982; Kennan and Walker, 2003; McCall and McCall, 1987).

Despite the fact the human capital and search theories are often regarded as competing, they nevertheless reach similar conclusions. First of all, they both predict that individuals with higher human capital are more likely to migrate. In the case of the human capital theory, this is due to the fact that individuals have to be compensated for their investment in education and normally have higher expected net migration benefits than less-educated people. In the case of job search, better-educated people need to be compensated for their higher reservation wage. However, one difference needs to be emphasised. In the human capital theory the migration propensity of each individual increases with education, while in the traditional job search theory, although on average higher human capital individuals are more mobile than lower human capital individuals, this does not necessarily hold true for every individual. In the search theory, whether or not an individual migrates is related to where the first acceptable job is located, that is, the job that meets the reservation wage. As jobs are randomly distributed over space and the process is sequential, it may be that some individuals are lucky enough to find an acceptable offer close to their current location. However, given that higher reservation wage jobs are expected to be more sparsely distributed in space, on average, higher human capital, and therefore higher reservation wage, individuals have to move further. As such, migration research, which combines human capital migration models (Sjaastad, 1962) with models of spatial job-search (Herzog et al., 1985, 1993), suggests that the likelihood of an individual moving between regions should be related to the individual's human capital, as well as the local economic and employment characteristics of both the region of origin and the destination region, plus a range of other personal characteristics. In particular, as well as human capital, current migration propensities are also found to be positively related to previous migration behaviour (DaVanzo and Morrison, 1981; DaVanzo, 1983) and also to being currently unemployed (DaVanzo, 1978), and negatively related to age (Schwartz, 1976; Inoki and Suruga, 1981; Bates and Bracken, 1987; Lundborg, 1991; Sandefur, 1985), with the exception of retirement migration. Other personal influences

such as gender (Faggian et al., 2007a) and ethnicity (Faggian et al., 2006) also play a role. In terms of regional influences, the effects of interregional unemployment differentials and interregional wages are unclear, with some observed flows being in the 'correct' direction (Cebula and Vedder, 1973; Rabianski, 1971; Rogers, 1967) while others appear to be rather perverse (Gordon and Molho, 1998; Hughes and McCormick, 1981, 1989, 1994; Jackman and Savouri, 1992; Millington, 2000; Pissarides and Wadsworth, 1989; Wall, 2001; Westerlund, 1997). The mixed results on wages and unemployment have led researchers to contemplate alternative explanations for interregional migration flows related to the compensating nature of wages, with respect to environmental variations and the regional differences in the quality of life (Graves, 1980; Porell, 1982; Schachter and Althaus, 1989, 1993; Clark and Cosgrove, 1991).

8.4 Human capital stocks, flows and mobility patterns

The analysis in the preceding sections generates two broad conclusions regarding the relationship between human capital and regional development. Firstly, from section 8.2 we see that aggregate stock of human capital is positively related to the level of economic growth and development. At the same time, from section 8.3 we see that individuals with higher human capital are more geographically and interregionally mobile. However, these observations alone are not sufficient to link human capital to regional development. The reason is that it depends on where such migrants move to.

In order to demonstrate how these various arguments are linked we must first clarify the analytical relationship between the models discussed in section 8.2 and section 8.3. The relationship between the aggregate analysis in section 8.2 and microeconomic analysis of section 8.3 is that the aggregate models described in section 8.2 treat human capital as a regional stock variable, and it is represented by the average human capital of the local labour multiplied by the number of such workers. On the other hand, while section 8.3 analyses the migration behaviour of individual people, the regional aggregation of these individual human capital migration movements produces a regional human capital flow variable. For growing regions this regional flow variable represents the increase in the stock of human capital per time period which is associated with migrant inflows into the host region. Conversely, for contracting regions, it represents the level of decrease in the human capital stock per time period which is associated with regional outflows of migrants in each time period. If the levels of interregional mobility are very low, or alternatively if the heterogeneity of migration propensities with respect to human capital is very low, then the human capital flow variables will be very low in comparison to the human capital stock variables. Regional growth will therefore be almost entirely dominated by internally generated local human capital, in a manner which is largely the same as for national growth models à la Lucas (1988). In terms of human capital-led growth, regions will therefore be largely closed free-standing autarkies. Under these conditions, any interregional migration of labour will generally represent an equilibrating process (Borts and Stein, 1964; Barro and Sala-i-Martin, 1990). Moreover, this outcome is even possible under conditions of highly heterogeneous migration propensities, as long as the geographical origin and destination patterns of individual migrations are randomly and equally distributed across all regions for all high human capital individuals. On the other hand, rather than being randomly and equally distributed, if the migration patterns of high human capital groups exhibit particular geographical polarities, in that they are

biased towards particular localities, then equilibrating processes associated with labour migration can be ruled out.

There are two possible scenarios here. The first scenario represents the case where net flows of human capital are biased in favour of particular destination regions, which benefit from net inflows of high human capital individuals from other source regions, regions which simultaneously exhibit net outflows of high human capital individuals. In general the destination region will exhibit a growing population while the origin regions will exhibit declining populations. In this situation, noticeably different migration propensities associated with individuals of different levels of human capital could easily lead to a growth process characterised by cumulative causation rather than by interregional convergence, as originally pointed out by Kaldor (1970). Such a cumulative process would be characterised by a situation whereby the more advanced higher-wage regions would benefit from the in-migration of workers in response to the higher wages, thereby increasing the effective internal regional demand, which in turn leads to greater localised knowledge investments and knowledge activities. Conversely, the poorer regions will experience a vicious circle characterised by the outmigration of workers, leading to a decrease in effective local demand, and a decline in knowledge investments and knowledge activities. This is the demand scale effect of human capital migration, and corresponds broadly to the home market effect in new economic geography. In addition to this scale effect there is also a labour composition effect, as workers who are most likely to move first to richer regions are those who are best trained. The result of this is that the regions which receive net inflows of labour also exhibit an upgrading of the average level of regional skills, while those that experience net outflows experience a downgrading of skills. As such, productivity in more advanced regions will grow while productivity in depressed regions will decline. The combination of these two effects associated with human capital migration engenders a cumulative growth process in the destination region and a cumulative decline in the origin region. As such, national growth may be associated with both regional growth and regional decline. The introduction of human capital can therefore imply increasing divergence at the regional level, because while more advanced regions benefit from a range of positive externalities, depressed regions will progressively suffer from outflows of skills. This represents a disequilibrium interregional labour migration adjustment mechanism, and one which can only be curtailed or reversed when the destination region exhibits binding capacity constraints. However, there is no reason why this cumulative process should be indefinite, as diseconomies of scale due to congestion and factor price appreciation may eventually emerge. Yet, anecdotal observations suggest that such processes may operate over very long time periods (Krugman, 1991).

The second scenario represents the case where particular types of high human capital individuals exhibit net inflows into particular localities so as to replace other particular types of human capital individuals who exhibit net outflows from the same region. In this situation, the absolute numbers of migrants moving in and out may not differ significantly, although the levels of human capital will consistently differ significantly. In this scenario, regions can settle into a permanent equilibrium characterised by disequilibrium. The example of this which is most often discussed is the case of cities such as London, which exhibit large gross in-migration and out-migration flows, but only small net inflows. Yet, these large migration flows represent a process of human capital churn

by which new skills, new ideas and new knowledge are continually being brought into the city. The logic of these migration patterns are related to intergenerational and life-cycle features, whereby migration plays a vital role in the social and career promotion processes of young people intent on increasing their human capital acquisition. In this schema, young people migrate to dominant knowledge cities such as London in order to enter the labour market, and subsequently rise within the urban and corporate employment hierarchies, only to move out later on in life as their preference structure evolves and when they can cash in on the capital gains from their properties and live a more comfortable life elsewhere. This phenomenon has come to be known as the 'escalator model' (Fielding, 1992a, 1992b, 1993; Faggian et al., 2007b), and it implies that dominant cities will be systematically characterised by inflows of young high human capital people and outflows of older high human capital people. Such escalator-type migration flows may be relatively stable in terms of little or no net inflows of people, but large gross flows of people both in and out of the city. Importantly, in this mechanism, the young university graduate immigrants not only represent high human capital but, following the Becker (1964) hypothesis, they are also at the age where their ability to learn and acquire human capital, and consequently to exhibit productivity gains, is also at its highest. Moreover, as well as being highly motivated to learn, their human capital is also associated with the newest vintages of technology. Finally, because of their early career stage, these high-quality and rapidly learning human capital workers are also relatively low-wage individuals. On the other hand, the outmigrants from these same regions tend to be higher-wage high human capital individuals whose rate of learning is low. As such, the dominant city destination region continually benefits in terms of efficiency wage effects.

Heterogeneous human capital migration flows can therefore contribute to redistributional effects between regions, in terms of their human capital stocks. Following the Lucas (1988) argument, any interregional redistribution of human capital stocks associated with these human capital flows can engender a variety of growth trajectories in different regions of the same country, some of which will mirror the national trend and some of which will differ markedly from the national growth trajectory. It depends on the particular geographical patterns of human capital mobility, and the ability of the individual region to retain, maintain or grow its stock of human capital.

From the above discussion the major feature which distinguishes regional growth processes from the aggregate national growth models is that the notion of endogeneity changes both subtly and fundamentally as we move from a national aggregate to a regional context. In national aggregate endogenous growth models the notion of endogeneity implies that a cumulative growth process is internally driven, where 'internal' here implies within the country. Moreover, such models tend to be framed within a largely closed-country framework, such as the aggregate US economy. On the other hand, in the above discussion of human capital interregional migration, growth in the destination region is not entirely endogenous, where 'endogenous' implies being internally generated, as implied in the models of Romer (1986) and Lucas (1988). This is because much of the regional growth stimulus is in the form of inflows of externally acquired human capital. The aspect of endogenous growth which is specifically internal to the host region is in the form of the local knowledge spillovers resulting from the interaction between the existing regional factors and the immigrant human capital. Conversely, for the origin region the process is in reverse. The aspect of endogenous growth which is specifically internal to the

origin region is the process of cumulative decline due to localised negative externalities. The outmigration of local human capital is driven by external stimuli.

This difference in terms of our definition of what is endogeneity, while appearing to be rather subtle, is in fact fundamental. When moving from a national aggregate context to a regional framework, we must also consider endogeneity in terms of the extent to which the region is open. As such, whereas aggregate growth models are basically intranational in construction, regional growth models are fundamentally interregional in construction. As such, it is perfectly correct to describe the system of interregional human capital migration as reflecting an endogenous growth process, whereas describing the growth process of individual cities or regions as being endogenous is not strictly correct, unless it is the form of a closed city framework. However, cities are notoriously open economies, with high levels of inward and outward mobility, so this case is generally not applicable.

8.5 Human capital, cities and skills: empirical evidence

Since the early 1990s, whilst slow economic convergence has been occurring between most countries, internal interregional economic divergence has been an increasing feature of almost all countries (Brakman and van Marrewijk, forthcoming), and there is increasing evidence to suggest that the role played by dominant cities has increased over recent decades (McCann, forthcoming). Urban areas are playing an ever increasing role in the global economy and this is as true for industrialised economies as it is for newly industrialising economies (Venables, 2006). However, as well as this scale effect, it appears that there is also a qualitative change in the role played by cities, which also accounts for this scale effect. The various theoretical arguments outlined in a range of papers (Gaspar and Glaeser, 1998; Storper and Venables, 2004; McCann, 2007, forthcoming) all imply that knowledge generation and acquisition has become even more localised over recent years, and empirical evidence supports this argument (Acs, 2002; Carlino et al., 2007). The contention here is that information and communication technologies and face-to-face contact have increasingly become mutual complements for each other, rather than substitutes (Gaspar and Glaeser, 1998). In the light of the human capital migration arguments discussed above, the implication of this is that cities which play the role of centres of knowledge exchange will increasingly be dominated by high human capital individuals, as increasingly mobile workers respond to the increasing wage premia associated with high value-added knowledge work in cities. Indeed, recent evidence suggests that not only is there an increasing share of university-educated human capital living and working in cities (Berry and Glaeser, 2005), but this proportion of university-educated workers is also correlated with the existing human capital stock (Berry and Glaeser, 2005), and both are correlated with the growth of the city (Glaeser et al., 1995; Glaeser and Shapiro, 2003; Berry and Glaeser, 2005; Shapiro, 2006). In the US there is no evidence of the levels of high school human capital playing any role whatsoever (Shapiro, 2006), and this corroborates the initial argument that it is tertiary-educated human capital which is now crucial for regional development. Additional evidence in support of the endogenous interregional human capital migration system described above also comes from the fact that after conditioning on individual characteristics it is clear that wages are indeed higher in high human capital cities (Shapiro, 2006). Furthermore, US cities are found to be becoming more dissimilar in terms of their human capital composition (Berry and Glaeser, 2005), such that regional divergence appears to have superseded previous decades of regional convergence (Berry and Glaeser, 2005).

It may well be the case that the USA is rather different to other countries, in that it is far more open regarding interregional mobility than many other countries. In response to exogenous employment shocks interregional migration is by far the most important adjustment mechanism in the USA (Blanchard and Kats, 1992), whereas for many parts of Europe (Decressin and Fátas, 1995; Broersma and van Dijk, 2002; Pekkala and Kangasharju, 2002; Gács and Huber, 2005) changes in local participation rates appear to be much more important. Some studies (Bils and Klenow, 2000) even doubt the positive relationship between human capital and economic growth discussed in section 8.2, suggesting that the causality here is weak and may also run the other way around, such that growth causes inflows of human capital. However, this counter-argument is also consistent with the human capital models discussed in section 8.3. The origin of the stimulus is not necessarily critical. What is important is whether there is a feedback mechanism between growth, human capital acquisition or attraction, and then growth again.

Previously, the evidence on this point was very limited because microeconomic data of interregional human capital flows were not available. However, very recent European evidence from the UK and Finland points exactly to these cumulative feedback processes. Since the late 1990s UK regions have been steadily diverging, and much of the evidence suggests that this divergence is associated with the mobility of human capital (HMT-DTI, 2001). In particular, there is strong evidence of an endogenous interregional migration process associated with the migration behaviour of graduate human capital. In the UK the most noticeable regional beneficiary here is that of the London economy. London benefits every year from a net in-migration of people aged 16–24. As a consequence, London has a relatively young workforce compared to the rest of the country (Oxford Economic Forecasting, 2004), and Dixon (2003) shows that London is a net recipient not only of young migrants, but also of migrants at higher educational levels. This escalator aspect of the London economy has had repercussions for other parts of the country. The regions immediately adjacent to London have benefited from human capital spillovers, whereas more geographically peripheral regions are increasingly suffering net outflows of human capital. Simultaneous equation techniques allow us to uncover these feedback mechanisms whereby regional knowledge outputs, in the form of patent applications, are systematically a function of graduate inflows of human capital from other regions, whereas there is no effect for human capital which is acquired locally (Faggian and McCann, 2006, 2009). The same finding is also evident for Finland. Using innovation survey data it becomes clear that locally acquired human capital plays no role in the innovation performance of Finnish high-technology firms, whereas human capital acquired from other regions is indeed a significant contributor of local knowledge outputs (McCann and Simonen, 2005; Simonen and McCann, forthcoming). Both of these sets of observations are consistent with the theoretical arguments in section 8.3 and also the empirical observation (Faggian and McCann, 2006; Faggian et al., 2007a and 2007b) that local labour inputs are generally of lower human capital than those sourced from other localities, because lower human capital is associated with lower interregional migration propensities.

The migration of human capital is obviously a complex mechanism in which different stimuli may dominate at different periods. While productivity growth appears to be intrinsically related to the migration behaviour of high-quality human capital, we also know that high human capital individuals are high wage earners. This adds an additional

complication to the issue, in that amenity goods are highly income-elastic goods. The more important is the role of human capital in growth, the more important also ought to be the role of amenity goods and quality-of-life issues in determining exactly where the human capital migrates to. The actual spatial distribution of productivity growth will therefore depend not only on the nature of human capital interactions, but also on the choices of where such high-quality human capital chooses to locate. The most recent evidence from US cities suggests that human capital accounts for 60 per cent of productivity growth while quality of life issues account for 40 per cent (Shapiro, 2006).

8.6 Some extensions and ambiguities of the concept of human capital

The specific definition of human capital that Becker (1964) originally employed was simply that of education and on-the-job training, and it remained so in the literature until the 1990s. However, the emergence of new growth theories in the late 1980s, in which both knowledge spillovers and human capital play a key role, led to a widening of the concept of human capital (see Faggian, 2005), as authors employed increasingly diverse sources of evidence and data for their analyses. The reason is that the minimalist Becker notion of human capital was still found to fall far short in terms of explanatory power when incorporated into the new growth models described above. Therefore, in the early 1990s the notion of human capital was first extended to include any natural or physical health and ability which improves an individual's acquisition of knowledge and skills. This was the first real twist on the original human capital concept. However, almost universal availability of healthcare in advanced economies mean that the additional explanatory power provided for by this widening of the concept at the regional level in advanced economies was limited.

A second twist to the original human capital argument, however, came from sociology and political science. It was also argued by sociologists that the learning environment in which human capital develops is itself also a socially constructed phenomenon (Putnam, 1993), depending on social norms. As such, under the broader concept of 'social capital', some scholars also began to include in their growth explanations any additional social or institutional features which they perceived as fostering individual learning and skills (Glaeser et al., 2002). Moreover, aspects of the concept of social capital have been subsequently extended in various ways with concepts such as network capital and institutional capital.

In terms of the human capital argument, however, the most important twist on the human capital concept comes from the concept of 'creative capital' which was popularised by Florida (2002, 2005a, 2005b). Although Florida prefers to use the terms 'creativity' or 'creative capital' rather than social capital, his work has much in common with that of Putnam (1993) in that its focus is on the role of social norms and values and the networks based on them. Unfortunately, the result of this widening and blurring of the concept of human capital has led to much confusion in the literature (McCann, 2007). There is nowadays no general consensus on the definition of human capital, which has simply come to mean any knowledge, skills and competencies embodied in individuals or their social relations that increase an individual's productivity. As such, the boundary between what is individual human capital and what is social capital has now become both very blurred and so broadly and loosely defined that it becomes extremely difficult to measure these concepts. Even in the high-profile examples of creative capital widely

discussed by Florida (2002), there is a problem of observational equivalence in that it is almost impossible to determine whether any of these issues are really actually anything different from the original and well-defined concept of human capital (Glaeser, 2004).

In terms of regional development, however, these are not trivial points. This is because on the basis of our above discussion regarding the notion of endogeneity at the regional level, it becomes clear that the human capital model and the creativity literature often imply quite different mechanisms and public policies choices. In particular, increasing human capital implies a greater tendency to mobility and migration and therefore less place-specificity, whereas the creativity literature emphasises increasing localisation and place-specificity. This dichotomy is also reflected in the equilibrium–disequilibrium debate regarding human capital migration, in that the equilibrium model gives primacy to the place characteristics whereas the disequilibrium model gives primacy to the economic characteristics. As such, the success of any public policy will therefore depend on which is the better-specified model in the circumstances.

While the creativity literature is currently more popular in policy circles, it is analytically much less powerful than the human capital argument (Glaeser, 2004). This has therefore led some economists (Solow, 2000; Arrow, 2000) to argue that the concepts should be completely abandoned or redefined. The reason is that none of the various extensions of the original human capital concept to any of the alternative definitions of social capital exhibit the time-preference investment characteristics inherent in physical or monetary capital, whereas human capital does (Becker, 1964). As such, the extended concepts appear to operate either as a substitute or a complement to market-based exchange and allocation (Stiglitz, 2000). Therefore, employing this type of terminology often causes confusion because it implies that social capital has production function characteristics similar to other production factors, whereas in fact that this cannot be assumed. As such, rather than using the term 'social capital', it would appear to be more appropriate from an analytical perspective simply to discuss these other notions of capital in terms of 'informal institutions' (Dasgupta, 2000), while leaving the original Becker (1964) definition of human capital intact. By splitting up our notions of capital in this manner we are still able to focus on the role played by human capital in regional development, and to identify the additional contribution played by other production factors and other issues.

8.7 Conclusions

Human capital and regional development is a complex topic. Regional models of human capital are both subtly and fundamentally different to national models because regions are open systems and the links between human capital and regional development are mediated via migration mechanisms. This chapter has charted the notion of human capital from its early non-spatial microeconomic usage and its subsequent aggregate non-spatial macroeconomic usage, through to the explicitly spatial application of the concept in labour migration models and then to our more recent understanding of the relationship between human capital migration and regional endogenous growth processes. Recent evidence from various countries suggests that the role and mobility of human capital is becoming more important as a determinant of regional performance, and that dominant cities in particular are increasingly the beneficiaries of these human capital mobility effects. In addition, the concept of human capital has itself also been broadened to take

account of additional influences. Finally, while the impact of human capital on regional development is determined by both economic drivers and amenity issues, at present the best evidence suggests that economic issues currently predominate.

Notes

1. See Brownrigg (1973), Lewis (1988), Bleaney et al. (1992), Armstrong (1993), McNicoll (1993), Dineen (1995), Harris (1997), Armstrong et al. (1997), Chatterton (1997), PWC (2001) and Wardle (2001).
2. Felsenstein addresses the issue of 'human capital long-term' effects, but just in terms of estimating the present value of future income stream for a cohort of graduates (1995–96). He estimates the present value to be around €31 million, but this only includes 'private' benefits accrued to the individuals, not the 'social' benefits, which should include also positive externalities to the region (Felsenstein, 1997).
3. Equation (8.1) can be adapted to take into account the distinction between pecuniary and non-pecuniary returns:
$$\int_0^T \left\{ e^{-r_\phi t} U[R_{\phi i}(t)] + e^{-r_\psi t} U[R_{\psi i}(t)] \right\} dt < \int_0^T \left\{ e^{-r_\phi t} U[R_{\phi j}(t)] + e^{-r_\psi t} U[R_{\psi j}(t)] \right\} dt - \int_0^T e^{-rt} [C_{ij}(t)] dt$$
where this time it is assumed that pecuniary and non-pecuniary returns are discounted by the individual at different discount rates (respectively r_ϕ and r_ψ), thus adding a greater level of generality to the model.
4. To name a few: Schwartz (1973), Schultz (1975) and DaVanzo (1983).
5. For more recent contributions on the subject see also McKenna (1990) and the very comprehensive review by Mortensen (1986).

References

Acs, Z.J. (2002), *Innovation and the Growth of Cities*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Armstrong, H.W. (1993), 'The local income and employment impact of Lancaster University', *Urban Studies*, **30** (10), 1653–68.

Armstrong, H.W., J. Darrall and R. Grove-White (1997), 'The local economic impact of construction projects in a small and relatively self-contained economy: the case of Lancaster University', *Local Economy*, August, 146–59.

Arrow, K.J. (2000), 'Observations on social capital', in P. Dasgupta and I. Serageldin (eds), *Social Capital: A Multifaceted Perspective*, Washington, DC: World Bank.

Barro, R.J. and X. Sala-i-Martin (1990), 'Economic growth and convergence across the United States', Working Paper 3419, NBER, Cambridge, MA.

Bartel, A. (1979), 'What role does job mobility play?', *American Economic Review*, **69**, 775–86.

Basker, E. (2003), 'Education, job search and migration', University of Missouri Working Paper, 02–16.

Bates, J. and I. Bracken (1987), 'Migration age profiles for local authorities areas in England, 1971–1981', *Environment and Planning A*, **19**, 521–35.

Becker, G. (1964), *Human Capital*, New York: NBER Columbia University Press.

Berry, C.R. and E.L. Glaeser (2005), 'The divergence of human capital levels across cities', *Papers in Regional Science*, **84**, 407–44.

Bils, M. and P.J. Klenow (2000), 'Does schooling cause growth?', *American Economic Review*, **90** (5), 1160–83.

Blanchard, O.-J. and L.F. Katz (1992), 'Regional evolutions', *Brookings Papers on Economic Activity*, **1**, 1–75.

Bleaney, M.F., M.R. Binks, D. Greenaway, G.V. Reed and D.K. Whynes (1992), 'What does a university add to its local economy?', *Applied Economics*, **24** (3), 305–11.

Bogue, D.J. (1985), *The Population of the United States*, New York: Macmillan.

Borts, G.H. and J.L. Stein (1964), *Economic Growth in a Free Market*, New York: Columbia University Press.

Brakman, S. and C. van Marrewijk (forthcoming), 'It's a big world after all: on the economic impact of location and distance', *Cambridge Journal of Regions, Economy and Society*.

Broersma, L. and J. van Dijk (2002), 'Regional labour market dynamics in the Netherlands', *Papers in Regional Science*, **81** (3), 343–64.

Brownrigg, M. (1973), 'The economic impact of a new university', *Scottish Journal of Political Economy*, **20** (2), 123–39.

Carlino, G.A., S. Chatterjee and R.M. Hunt (2007), 'Urban density and the rate of invention', *Journal of Urban Economics*, **61** (3), 389–419.

Cebula, R.J. and R.K. Vedder (1973), 'A note on migration, economic opportunity, and quality of life', *Journal of Regional Science*, **13** (2), 205–11.

Chatterton, P. (1997), 'The economic impact of the University of Bristol on its region', www.bris.ac.uk/Publications/Chatter/impact.htm.

Clark, D.E. and J.C. Cosgrove (1991), 'Amenities versus labor market opportunities: choosing the optimal distance to move', *Journal of Regional Science*, **31** (3), 311–28.
Dasgupta, P. (2000), 'Economic progress and the idea of social capital', in P. Dasgupta and I. Serageldin (eds), *Social Capital: A Multifaceted Perspective*, Washington, DC: World Bank, pp. 325–424.
DaVanzo, J. (1976), 'Differences between return and non-return migration: an econometric analysis', *International Migration Review*, **10** (1), 13–27.
DaVanzo, J. (1978), 'Does unemployment affect migration? Evidence from micro data', *Review of Economics and Statistics*, **60**, 504–14.
DaVanzo, J. (1983), 'Repeat migration in the United States: who moves back and who moves on?', *Review of Economics and Statistics*, **65**, 552–9.
DaVanzo, J. and P.A. Morrison (1981), 'Return and other sequences of migration in the United States', *Demography*, **18** (1), 85–101.
Decressin, J. and A. Fátas (1995), 'Regional labour market dynamics in Europe', *European Economic Review*, **39**, 1627–55.
Dineen, D.A. (1995), 'The role of a university in regional economic development: a case study of the University of Limerick', *Industry and Higher Education*, **9**, 140–48.
Dixon, S. (2003), 'Migration within Britain for job reasons', *Labour Market Trends*, April, 191–201.
Faggian, A. (2005), 'Human capital', in R. Caves (ed.), *Encyclopaedia of the City*, New York: Routledge.
Faggian, A. and P. McCann (2006), 'Human capital flows and regional knowledge assets: a simultaneous equation approach', *Oxford Economic Papers*, **52**, 475–500.
Faggian, A. and P. McCann (2009), 'Human capital, graduate migration and innovation in British regions', *Cambridge Journal of Economics*, forthcoming.
Faggian, A., P. McCann and S. Sheppard (2006), 'An analysis of ethnic differences in UK graduate migration behaviour', *Annals of Regional Science*, **40** (2), 461–71.
Faggian, A., P. McCann and S. Sheppard (2007a), 'Some evidence that women are more mobile than men: gender differences in UK graduate migration behaviour', *Journal of Regional Science*, **47** (3), 517–39.
Faggian, A., P. McCann and S. Sheppard (2007b), 'Human capital, higher education and graduate migration: an analysis of Scottish and Welsh students', *Urban Studies*, **44** (13), 2511–28.
Fahr, R. and U. Sunde (2002), 'Employment status, endogenous regional mobility, and spatial dependencies in labor markets', IZA Discussion Paper, No. 521.
Felsenstein, D. (1997), 'Estimating some of the impacts on local and regional economic development associated with Ben-Gurion University of the Negev', Working Paper, Negev Center for Regional Development, Ben-Gurion University of the Negev.
Fielding, A.J. (1992a), 'Migration and social mobility: South East England as an escalator region', *Regional Studies*, **26** (1), 1–15.
Fielding, A.J (1992b), 'Migration and social change', in J. Stillwell, P. Rees and P. Boden (eds), *Migration Processes and Patterns. Volume 2: Population Redistribution in the United Kingdom*, London and New York: Belhaven Press, pp. 225–47.
Fielding, A.J. (1993), 'Migration and the metropolis: recent research on the causes and consequences of migration to the Southeast of England', *Progress in Human Geography*, **17** (2), 195–212.
Florida, R. (2002), *The Rise of the Creative Class*, New York: Basic Books.
Florida, R. (2005a), *The Flight of the Creative Class: The New Global Competition for Talent*, New York: Harper Collins.
Florida, R. (2005b), *Cities and The Creative Class*, London: Routledge.
Friedman, M. and S. Kuznets (1954), *Income from Independent Professional Practice*, Princeton, NJ: Princeton University Press.
Gács, V. and P. Huber (2005), 'Quantity adjustments in the regional labour markets of EU candidate countries', *Papers in Regional Science*, **84** (4), 553–74.
Gaspar, J. and E.L. Glaeser (1998), 'Information technology and the future of cities', *Journal of Urban Economics*, **43**, 136–56.
Glaeser, E.L. (2004), 'Review of Richard Florida's *The Rise of the Creative Class*', mimeo, Department of Economics, Harvard University.
Glaeser, E.L. and J.M. Shapiro (2003), 'Urban growth in the 1990s: is city living back?', *Journal of Regional Science*, **43** (1), 139–65.
Glaeser, E.L., D.I. Laibson and B. Sacerdote (2002), 'An economic approach to social capital', *Economic Journal*, **112** (483), 437–58.
Glaeser, E.L., J.A. Scheinkman and A. Shleifer (1995), 'Economic growth in a cross section of cities', *Journal of Monetary Economics*, **36**, 117–43.
Gordon, I. and I. Molho (1998), 'A multi-stream analysis of the changing pattern of inter-regional migration in Great Britain, 1960–91', *Regional Studies*, **32** (4), 309–23.

Gordon, I. and R. Vickerman (1982), 'Opportunity, preference and constraint: an approach to the analysis of metropolitan migration', *Urban Studies*, **19** (3), 247–61.
Graves, P.E. (1980), 'Migration and climate', *Journal of Regional Science*, **20** (2), 227–37.
Harris, R.I.D. (1997), 'The impact of the University of Portsmouth on the local economy', *Urban Studies*, **34** (4), 605–26.
Hart, R.A. (1975), 'Interregional economic migration: some theoretical considerations (Part II)', *Journal of Regional Science*, **15** (3), 289–305.
Herzog, H.W., R.A. Hofler and A.M. Schlottmann (1985), 'Life on the frontier: migrant information, earnings and past mobility', *Review of Economics and Statistics*, **67**, 373–82.
Herzog, H.W., A.M. Schlottmann and T.P. Boehm (1993), 'Migration as spatial job-search: a survey of empirical findings', *Regional Studies*, **27** (4), 327–40.
HMT-DTI (2001), *Productivity in the UK: 3 The Regional Dimension*, London: H.M. Treasury and Department for Trade and Industry.
Hughes, G.A. and B. McCormick (1981), 'Do council housing policies reduce migration between regions?', *Economic Journal*, **91**, 919–37.
Hughes, G.A. and B. McCormick (1989), 'Does migration reduce differentials in regional unemployment rates?', in J. van Dijk, H. Folmer, H.W. Herzog and A.M. Schlottman (eds), *Migration and Labour Adjustment*, Boston, MA and London: Kluwer.
Hughes, G.A. and B. McCormick (1994), 'Did migration in the 1980s narrow the North–South divide?', *Economica*, **61** (244), 509–27.
Inoki, T. and T. Suruga (1981), 'Migration, age and education: a cross-sectional analysis of geographic labor mobility in Japan', *Journal of Regional Science*, **21** (4), 507–17.
Jackman, R. and S. Savouri (1992), 'Regional migration in Britain: an analysis of gross flows using NHS Central Register data', *Economic Journal*, **102** (415), 1433–50.
Kaldor, N. (1970), 'The case for regional policies', *Scottish Journal of Political Economy*, **17**, 337–47.
Kennan, J. and J.R. Walker (2003), 'The effect of expected income on individual migration decisions', NBER Working Paper, No. 9585.
Krugman, P. (1991), *Geography and Trade*, Cambridge, MA: MIT Press.
Levy, M. and W. Wadycki (1974), 'Education and the decision to migrate: an econometric analysis of migration in Venezuela', *Econometrica*, **42** (2), 377–88.
Lewis, J.A. (1988), 'Assessing the effect of the Polytechnic, Wolverhampton on the local community', *Urban Studies*, **25**, 53–61.
Lippman, S.A. and J.J. McCall (1976a), 'The economics of job search: a survey (Part I)', *Economic Inquiry*, **14**, 155–89.
Lippman, S.A. and J.J. McCall (1976b), 'The economics of job search: a survey (Part II)' *Economic Inquiry*, **14**, 347–68.
Lippman, S.A. and J.J. McCall (1979), *Studies in the Economics of Search*, Amsterdam: North-Holland.
Lucas, R. (1988), 'On the mechanism of economic development', *Journal of Monetary Economics*, **22**, 3–42.
Lundborg, P. (1991), 'Determinants of migration in the Nordic labor market', *Scandinavian Journal of Economics*, **93** (3), 363–76.
McCall, B.P. and J.J. McCall (1987), 'A sequential study of migration and job search', *Journal of Labor Economics*, **5**, 452–76.
McCann, P. (2007), 'Sketching out a model of innovation, face-to-face interaction and economic geography', *Spatial Economic Analysis*, **2** (2), 117–34.
McCann, P. (forthcoming), 'Globalization and economic geography: the world is curved, not flat', *Cambridge Journal of Regions, Economy and Society*.
McCann, P. and J. Simonen (2005), 'Innovation, knowledge spillovers and local labour markets', *Papers in Regional Science*, **84** (3), 465–85.
McKenna, C.J. (1990), 'The theory of search in labour markets', in D. Sapsford and Z. Tzannatos (eds), *Current Issues in Labour Economics*, Basingstoke: Macmillan, pp. 33–62.
McNicoll, I.H. (1993), 'The impact of Strathclyde University on the economy of Scotland', Department of Economics, University of Strathclyde.
Millington, J. (2000), 'Migration and age: the effect of age on sensitivity to migration stimuli', *Regional Studies*, **34** (6), 521–33.
Mincer, J. (1958), 'Investment in human capital and personal income distribution', *Journal of Political Economy*, **66**, 281–302.
Mincer, J. (1974), *Schooling, Experience and Earnings*, New York: Columbia University Press.
Mortensen, D.T. (1986), 'Job search and labor market analysis', in O. Ashenfelter and R. Layard (eds) *Handbook of Labor Economics*, Amsterdam: North-Holland.
Nijkamp, P. and J. Poot (1998), 'Spatial perspectives on new theories economic growth', *Annals of Regional Science*, **32**, 7–27.
OECD (2006), *Education at a Glance: OECD Indicators 2006*, Paris: OECD.
Oxford Economic Forecasting (2004), *London's Linkages with the Rest of the U.K.*, Corporation of London.

Pekkala, S. and A. Kangasharju (2002), 'Regional labor markets in Finland: adjustment to total versus region-specific shocks', *Papers in Regional Science*, **81** (3), 329–42.
Pissarides, C.A. (1976), *Labour Market Adjustment*, Cambridge: Cambridge University Press.
Pissarides, C.A. and J. Wadsworth (1989), 'Unemployment and the inter-regional mobility of labour', *Economic Journal*, **99**, 739–55.
Plane, D. and P. Rogerson (1994), *The Geographical Analysis of Population: With Applications to Planning and Business*, New York: John Wiley & Sons.
Polachek, S. and F. Horvath (1977), 'A life cycle approach to migration: analysis of the perspicacious peregrinator', in R. Ehrenberg (ed.), *Research in Labor Economics*, Greenwich, CT: JAI Press.
Polachek, S. and S. Siebert (1993), *The Economics of Earnings*, Cambridge: Cambridge University Press.
Porell, F.W. (1982), 'Intermetropolitan migration and quality of life', *Journal of Regional Science*, **22** (2), 137–58.
PWC, (2001), 'University of Waterloo – regional economic benefits study', PricewaterhouseCoopers, Ontario.
Putnam, R.D. (1993), *Making Democracy Work: Civic Traditions in Modern Italy*, Princeton, NJ: Princeton University Press.
Rabianski, J. (1971), 'Real earnings and human migration', *Journal of Human Resources*, **6**, 185–92.
Rebhun, U. (2003), 'The changing roles of human capital, state context of residence, and ethnic bonds in interstate migration: American Jews 1970–1990', *International Journal of Population Geography*, **9**, 3–21.
Rogers, A. (1967), 'A regression analysis of interregional migration in California', *Review of Economics and Statistics*, **49** (2), 262–67.
Rogerson, P. (1982), 'Spatial models of search', *Geographical Analysis*, **14**, 217–28.
Romer, P.M. (1986), 'Increasing returns and long-run growth', *Journal of Political Economy*, **94**, 1002–37.
Romer, P.M. (1990), 'Endogenous technological change', *Journal of Political Economy*, **98**, 71–102.
Romer, P.M. (1994), 'The origins of endogenous growth', *Journal of Economic Perspectives*, **8**, 3–22.
Sandefur, G.D. (1985), 'Variations in interstate migration of men across the early stages of the life cycle', *Demography*, **22** (3), 353–66.
Schachter, J. and P.G. Althaus (1989), 'An equilibrium model of gross migration', *Journal of Regional Science*, **29** (2), 143–59.
Schachter, J. and P.G. Althaus (1993), 'The assumption of equilibrium in models of migration', *Journal of Regional Science*, **33** (1), 85–8.
Schultz, T. (1975), 'The value of the ability to deal with disequilibria', *Journal of Economic Literature*, **13** (3), 827–46.
Schwartz, A. (1973), 'Interpreting the effect of distance on migration', *Journal of Political Economy*, **81**, 1153–69.
Schwartz, A. (1976), 'Migration, age and education', *Journal of Political Economy*, **24**, 701–20.
Shapiro, J.M. (2006), 'Smart cities: quality of life, productivity, and the growth effects of human capital', *Review of Economics and Statistics*, **88** (2), 324–35.
Simonen, J. and P. McCann (forthcoming), 'Firm innovation: the influence of R&D cooperation and the geography of human capital inputs', *Journal of Urban Economics*.
Sjaastad, L. (1962), 'The costs and returns of human migration', *Journal of Political Economy*, **70** (5), 80–93.
Solow, R.M. (1956), 'A contribution to the theory of economic growth', *Quarterly Journal of Economics*, **70**, 65–94.
Solow, R.M. (2000), 'Notes on social capital and economic performance', in P. Dasgupta, and I. Serageldin (eds), *Social Capital: A Multifaceted Perspective*, Washington, DC: World Bank.
Stark, O. and J. Taylor (1991), 'Migration incentives, migration types: the role of relative deprivation', *Economic Journal*, **101** (408), 1163–78.
Stiglitz, J.E. (2000), 'Formal and informal institutions', in P. Dasgupta and I. Serageldin (eds), *Social Capital: A Multifaceted Perspective*, Washington, DC: World Bank.
Storper, M. and A.J. Venables (2004), 'Buzz: face-to-face contact and the urban economy', *Journal of Economic Geography*, **4**, 351–70.
Swan, T. (1956), 'Economic growth and capital accumulation', *Economic Record*, **32**, 334–61.
Topel, R.H. (1986), 'Local labor markets', *Journal of Political Economy*, **94** (3), 111–43.
Venables, A.J. (2006), 'Shifts in economic geography and their causes', *Federal Reserve Bank of Kansas City Economic Review*, **91** (4), 61–85.
Wall, H.J. (2001), 'Voting with your feet in the United Kingdom: using cross-migration rates to estimate relative living standards', *Papers in Regional Science*, **80**, 1–23.
Wardle, P. (2001), 'The economic impact of four large education institutions on the Canterbury district economy', Canterbury City Council.
Westerlund, O. (1997), 'Employment opportunities: wages and interregional migration in Sweden', *Journal of Regional Science*, **37** (1), 55–73.
Yezer, A. and L. Thurston (1976), 'Migration patterns and income change: implications for the human capital approach to migration', *Southern Economic Journal*, **42**, 693–702.

9 Infrastructure and regional development

Johannes Bröcker and Piet Rietveld

9.1 Introduction

Regional development is the result of mutually related decisions made by private and public actors. In the present chapter we will focus on one particular type of decision: the provision of infrastructure, mainly by the public sector, but also possibilities for private supply will be explored. The impacts will be measured in particular in terms of productivity and welfare. The main type of infrastructure to be studied here is transport infrastructure, but many of the results will also apply to other infrastructure types.

Regional impacts of infrastructure supply are of interest for two reasons. First, infrastructure investment plans are often motivated by regional policy goals. They are intended to benefit lagging regions. Hence, assigning benefits to regions is vital in this context. Second, assessing benefits by regions is needed for assigning the planning and decision responsibility as well as the financial burden in a proper way. Local jurisdictions should decide upon projects not having significant spillovers to other jurisdictions, and they should fully pay for them. In case of spillovers, decentralized solutions are still possible through negotiations of jurisdictions, as advocated by Coase. In this case local decision-makers should at least have a rough idea about who gains what, in order to attain an agreement about projects that generate enough benefit to make the citizens of all jurisdictions involved better off. Transaction costs make a decentralized agreement impracticable, however, if too many jurisdictions are involved. Either decision-making and financing have to be raised to a higher administrative level in this case, or proper incentives for decentralized decisions have to be generated by matching grants compensating for spillovers. In any case, an assessment of spillovers is needed for designing an appropriate institutional arrangement.

It appears that there is often considerable uncertainty on the regional economic effects of infrastructure supply. For example, the literature on the general productivity effects of infrastructure triggered by Aschauer (1989) has led to widely varying outcomes. And in the case of planning specific elements in infrastructure networks there is also often considerable debate on the effects, including the 'indirect effects' or wider economic benefits that are beyond the welfare effects measured within the transport system itself. In this review chapter we will focus on the spatial distribution of infrastructure impacts. This does not cover the whole field of infrastructure research. For example, tremendous research effort has gone into quantifying monetary values of time-saving, accident injuries and fatalities and environmental impacts on the local (noise, toxic emissions) and global (biodiversity, climate) scale. This is for good reasons. Time savings, accidents and environment impacts are the main non-monetary items to which money values have to be assigned in cost–benefit analysis; and many of the methods applied are still unsatisfying. But in order to achieve sufficient focus, we will not contribute to these issues here. We just assume environmental impacts and safety issues away, and we take for granted that generalized costs of trips, covering out-of-pocket as well as time costs, are perfectly known.

We also assume project costs to be known, and we totally disregard the time dimension by measuring project costs in annuities, assuming that all the subtle questions of choosing depreciation and discount rates have already been solved.

The structure of this chapter is as follows. We start with a discussion of definitions and measurements of infrastructure (section 9.2). A general discussion of infrastructure impacts is given in section 9.3. Section 9.4 reviews the literature on productivity effects of infrastructure and discusses important themes such as the specification of services provided with infrastructure, spatial spillovers, causality issues and crowding-out effects. In the second part of the contribution we shift the focus from a productivity orientation to a welfare orientation. We first review the fundamentals of the theory of optimal provision of infrastructure in section 9.5. Section 9.6 briefly introduces the most advanced technique available for assessing regional welfare effects of infrastructure, namely spatial computable general equilibrium (CGE) analysis. Section 9.7 then more extensively develops a method that is much less demanding in terms of data as well as computational complexity, but still theoretically well founded and closely related to a familiar approach in regional science: gravity analysis. Section 9.8 outlines further thoughts on 'wider economic effects', that is effects that are not accounted for by the surplus measures considered in sections 9.5 and 9.7. Section 9.9 summarizes the main findings.

9.2 Defining and measuring infrastructure

The literature on infrastructure impacts on the economy is characterized by a rather pluriform approach to defining or delimiting it. For example, Canning (1998) in his effort to develop a worldwide database of infrastructure covers the following components:

- telephones and telephone main lines;
- electricity generating capacity;
- roads (length in kilometres; paved, non-paved);
- railways (length in kilometres).

In an earlier study Biehl (1986, 1993) developed a database for regions within the EU covering the same elements as mentioned by Canning (communications, energy supply and transport) but at a more detailed level, for example distinguishing for transport: roads, rail, waterways, airports, seaports and pipelines. Further, he adds infrastructure components related to water management, environmental management, education, health service provision, sports and tourist facilities, social infrastructure, cultural facilities and natural endowments. Both for the Biehl and the Canning databases, data availability played a large role in determining the delimitation of the infrastructure concept.

A broader discussion of delimitations and definitions of infrastructure found in the literature is given in Rietveld and Bruinsma (1998). Many contributions to the field mention the public good character of infrastructure, involving the notions of non-rivalness and non-excludability. However, it is not difficult to see that important parts of what is commonly considered as infrastructure (for example rail and airports) involve services that are excludable (users can be forced to pay for the services they consume) and/or rival (congestion is a relevant theme). Thus, in a strict sense, only a small part of what is commonly understood by infrastructure really is a public good.

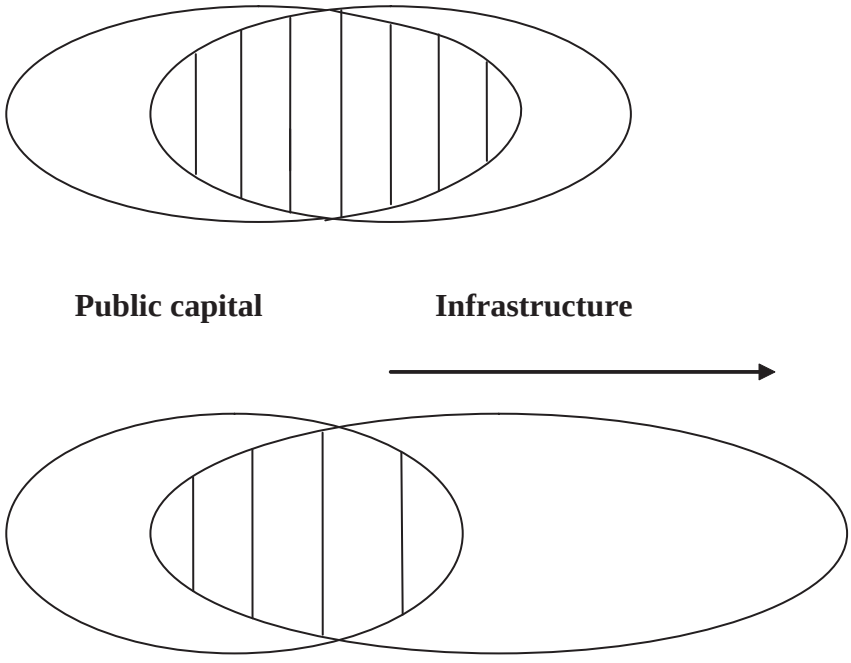


Figure 9.1 Decreasing overlap between public capital and infrastructure

A related definition is that infrastructure is capital that is publicly provided, or where the public sector at least has a large role in service provision. However, the public sector role varies among countries, and technological developments may have strong impacts on the role of the public sector. For example, telecommunication has in most countries shifted from the public sector to the private sector. Further, information and communication technology (ICT) developments make it much easier nowadays to let individuals pay for the use of transport networks, implying increasing opportunities for private sector involvement. Note also, that within the transport sector pipelines have always been supplied predominantly by the private sector. Thus, one can observe a limited and even decreasing overlap between what is commonly understood by infrastructure, and publicly provided capital (see Figure 9.1).

A common approach to measuring infrastructure stocks is to base them on the values of infrastructure investments measured in monetary terms during a certain period. This requires the use of the well-known perpetual inventory methods, involving the use of expected service lives and rates of deterioration. Since countries may differ in their use of these parameters, this is a potential source of incomparability of data from different countries. A more serious point, however, is that countries may vary widely in terms of the per-unit costs of the infrastructure, given that input prices may be very different, and this also holds for public sector efficiency (Canning, 1998). A third factor concerns natural conditions that may vary among countries, but also within countries. Obvious examples are the high costs of bridges and tunnels needed in mountain areas or for water crossings. Even apart from these factors, highway and rail construction in urban areas will be much more expensive than in rural areas given the differences in land prices, number of crossing

Table 9.1 Infrastructure stock, infrastructure services and quality aspects, some examples

Infrastructure stock	Infrastructure services	Quality aspects
Expressway (in kms)	Vehicle movements per day, vehiclekms per day, accessibility	Maximum possible speed, average speed, maximum axle load, available during X hours per year, accidents, variation in travel times
Railway (kms)	Frequency of trains Seatkms	Maximum possible speed, average speed, availability, accidents, variation in travel times
Electricity supply capacity (MW)	KWh	Availability during the year, loss of electricity in the system

infrastructure links, and cost increases due to the extra need to reduce the environmental burden in urban areas. Thus in two countries that happen to have an identical infrastructure stock in physical terms, the valuation in monetary terms may be very different.

Although one might expect a preference for physical measures of infrastructure, it appears that a large part of the research on infrastructure impacts on the economy is based on monetary values of stocks. One reason for this is that monetary measures are convenient when adding infrastructure elements of various types. Another reason is probably that a considerable part of the research in this field has a macroeconomic orientation, implying a tendency to use infrastructure data in monetary terms instead of physical measures. Also the availability of physical measures of the infrastructure stock may be problematic.

One of the reasons why physical measures of the infrastructure stock make more sense than monetary measures is that they provide a natural starting point for the measurement of infrastructure services like vehiclekms, and so on, which are the result of combining the infrastructure stock with other forms of capital, like rolling stock in the case of transport infrastructure (see Table 9.1). Further, the concept of accessibility, being an indicator of the potential of interaction provided by an infrastructure network linking nodes with different features may be considered as an indicator of potential services generated. The accessibility concept will be discussed in more detail in section 9.4.

A final point addressed in Table 9.1 concerns the quality of infrastructure services measured in terms of indicators like speed, reliability, availability and safety. This is a theme with considerable scientific and policy interest. From a policy perspective it is important to find the appropriate balance between construction of new infrastructure and maintaining and upgrading existing stock (see for example Briceno et al., 2004). Indeed, many industrialized countries now experience a regime shift away from adding new infrastructure towards maintaining existing infrastructure and improving quality of services of existing infrastructure. In some cases such as congestion in transport networks, the quality dimension has received ample attention in the scientific arena, but the link with the current literature on productivity impacts of infrastructure has remained mainly implicit as we will see.

Canning (1998) indicates that in the infrastructure domain, data on quantities are better than those on quality. Therefore, customer satisfaction is sometimes used as an alternative way of measuring quality. Table 9.2 summarizes some findings on subjective assessments of quality among various infrastructure types and types of countries.

Table 9.2 Commercial users' views on the quality of infrastructure services, by country income group (2000–02)

Income group	Electricity	Telecoms	Roads	Railroads	Ports	Airports
Low	2.6	3.4	3.4	2.7	2.6	3.6
	(9)	(9)	(27)	(9)	(9)	(9)
Lower middle	4.2	4.9	4.2	2.6	3.5	4.2
	(25)	(25)	(24)	(25)	(25)	(25)
Upper middle	5.1	5.6	4.1	2.9	3.8	4.5
	(20)	(20)	(18)	(26)	(20)	(20)

Note: Ratings are on a scale of 1 to 7, with 7 indicating the highest quality. Figures in parentheses indicate the number of countries for which data are available.

Source: Briceno et al. (2004).

A striking feature of Table 9.2 is that there is an almost monotone relationship between income level in a country and the degree of user satisfaction (valued by actors of commercial and industrial communities). However, rail is hardly participating in this trend. A second observation is that satisfaction levels remain far from the best-practice scores of 7, implying that quality is a problem everywhere, even in upper-middle-income countries.

9.3 Infrastructure impacts

Figure 9.2 illustrates some main effects of infrastructure investment on production as measured by gross domestic product (GDP). The short-term effects are shown in the left side of the figure. Infrastructure investment leads to expanded activity in the construction sector as well as in the construction materials sector. These short-term effects can be analysed by means of input–output analysis. At the lower part of the left side are the fiscal effects of this expenditure, related to the way the public sector collects the resources needed for the investments. At the right side of the figure we have the long-run effects of the infrastructure. These effects consist of two parts: productivity effects that can be economy-wide, and effects in the maintenance sector. Both will probably increase in the course of time. First, the productivity effects will take time to materialize due to inertia in the economy. Second, when there is a growing economy and the infrastructure is subject to congestion effects, the effects of not implementing the project will increase in the course of time. Third, maintenance costs would increase with the intensity of use. Note that where in a picture like this all periods receive equal weight, in most economic analyses future effects will typically be discounted, implying a relatively strong role for the effects in the short term.

In terms of uncertainty about these effects, different mechanisms are at stake. First, short-term effects are intrinsically easier to estimate than long-run effects. Nevertheless, as emphasized by Flyvbjerg (2003), there are tendencies within political decision-making processes that the construction costs are strategically underestimated by public actors that are committed to the realization of certain projects. A related reason for cost increases is that local actors fearing negative externalities of projects only cooperate when expensive preventive measures are taken, or when financial compensation is given. This of course leads to biases in the estimation of economic effects during the construction phase. For

the positive long-run effects there may be similar tendencies that these are systematically overestimated (see Flyvbjerg et al., 2005). Indeed, in the present chapter we will find that there is considerable uncertainty about the actual contribution of infrastructure to the economy. Two main sources of this uncertainty are: model uncertainty, related to imprecise knowledge on essential parameters and the fact that infrastructure impacts depend on the state of the economy, which is by definition difficult to predict for a long-run period. In particular the effects of infrastructure projects that are meant to reduce congestion will depend critically on the levels of congestion during the period after the completion of the project, and these levels of course depend on the state of the economy. For a proper assessment of infrastructure projects, broader network conditions should be considered. In particular the initial conditions in the infrastructure network are relevant in this respect. When an already well-developed network exists, the effects of network extensions will be smaller than when a network is non-existent, when it is highly congested or when a project concerns a missing link (Rietveld and Bruinsma, 1998; Fernald, 1999). A related consideration is that in many cases an infrastructure project consists of several components. For example, building a high-speed rail connection between two cities is often accompanied by the reconstruction of railway stations and the improvement of local accompanying measures such as the provision of adequate feeder services. The final effect on transport and land use should therefore not only be attributed to the construction of the railway line, but also to the other components.

Of special interest is the spatial dimension of infrastructure impacts. The region where the infrastructure is built will probably experience most of the production effects in the construction sector, although this depends on the type of infrastructure. Advanced projects may lead to the need to involve specialized companies and workers from other regions. Similar leakages will occur with the construction materials sector. When a regional project is carried out by means of national funds, the negative fiscal effects materialize at the national level, creating a clear lack of balance between who bears the favourable and unfavourable effects on production. Also for the long run, spatial dimensions are important. When road or rail infrastructure is built to accommodate long-distance transport flows, the regions affected by negative externalities may not be the ones where the main positive production effects take place. This leads to the important theme of spatial spillovers in infrastructure research, a theme to which we will return in a next section.

Figure 9.2 does not give the full picture of infrastructure-related effects on the economy. First of all, it ignores broader welfare effects, for example time gains in passenger transport that, apart from business travel, would not be traced in the national accounts on which the production changes are based. Depending on the type of infrastructure, the importance of the effects that do not show up in the national accounts will vary. In the case of road investments in industrialized countries, it is common that the welfare effects in reductions of travel time measured by means of the value of time (Small, 1992) dominate the productivity effects measured via GDP (Mohring, 1975). Welfare effects will be addressed in section 9.5.

Further, to keep the analysis focused, the environmental effects of the construction, existence and use of the infrastructure have been ignored in this chapter (for reviews on the environmental effects of transport see, for example, Rietveld, 2005; Stead, 2007). Hence there is only a limited connection between productivity studies and cost–benefit studies carried out for specific infrastructure projects.

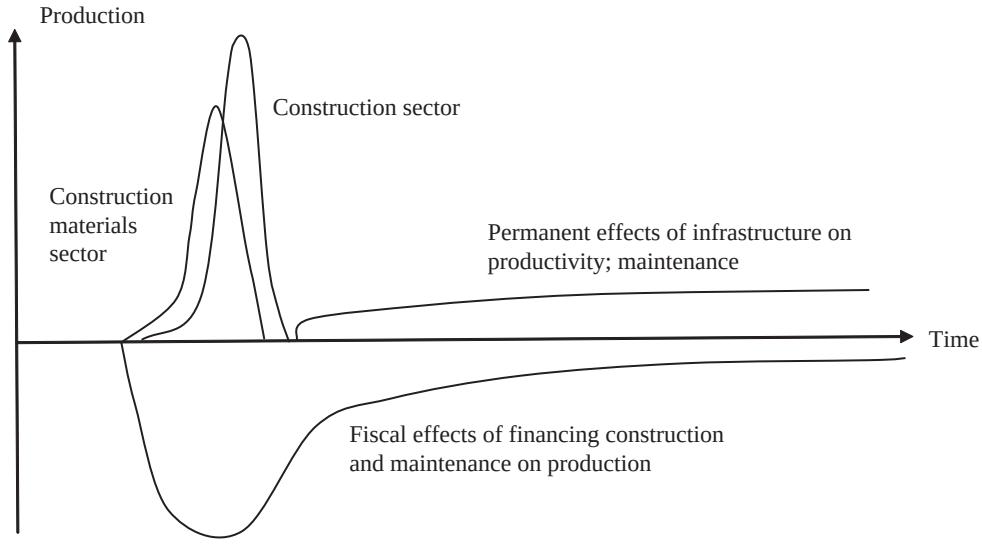


Figure 9.2 Short-run and long-run effects of infrastructure investments

Research on infrastructure impacts is part of a broader field of the effect of government policies on long-run growth and productivity (Easterly and Levine, 2001; Nijkamp and Poot, 2004). For example, Nijkamp and Poot distinguish different areas of fiscal policies (that are sometimes partly overlapping): investment in infrastructure, and investments in education, defence and government expenditure at large (but excluding transfers). Table 9.3 summarizes the results of a set of 113 studies that appeared in main economic journals during the past decades. The focus was on the sign of the impact of public sector policy on productivity and growth. It appears that infrastructure and education are two fields where a positive and significant sign is found in the majority of the cases, whereas with defence expenditure and total government expenditure (sometimes referred to as government size) the negative outcomes are dominating. Although the majority of studies find a positive effect for infrastructure, there is considerable uncertainty on its size. The background of this will be discussed in section 9.4, where we will also address the adverse effects of government spending on production values related to the crowding-out problem of government expenditure.

9.4 Infrastructure as a production factor

The work of Aschauer (1989) meant the start of an intensive scientific and political debate on the contribution of infrastructure to productivity. This debate mainly took place within the domain of macroeconomics, leading to the use of production functions in macroeconomic terms, later followed by cost function approaches, as the main vehicle of research. It is interesting to note that a considerable time before the end of the 1980s similar contributions appeared in the field of regional economics (for example, Mera, 1973; Fukuchi, 1978; Blum, 1982; Costa et al., 1987) but these did not trigger much attention. Clearly, the state of the world economy at that particular time, and in particular that of the USA with its long-run decline in productivity growth, has played a large role here.

Table 9.3 Public sector expenditure and its effect on national production values according to studies that appeared in articles in refereed economic journals

Type of fiscal policy	Number of studies considered	Proportion concluding positive impact	Proportion concluding negative impact	Proportion concluding inconclusive impact
Education	12	0.92	0.00	0.08
Infrastructure	39	0.72	0.08	0.20
Defence	21	0.05	0.52	0.43
Total government consumption ('government size')	41	0.17	0.29	0.54
All types	113	0.41	0.23	0.35

Source: Nijkamp and Poot (2004).

The basic starting point of the analysis is a production function where in addition to the standard production factors of labour and capital, a distinction is made between private capital K, and public capital G. In a time series context the production function with production level Y can then be formulated as:

$$Y_t = f(L_t, K_t, G_t),$$

where t indicates the year, and the use of the Cobb–Douglas production function is the standard one. Aschauer found a very high elasticity for the contribution of public capital to production of 0.39 for a time series of national data in the USA, implying a strong case for the public sector to increase its investment since it implies a very high rate of return on investment in public capital. The implied increase in production as a consequence of an investment in public capital ($\Delta Y/\Delta G$) would equal (0.4) (Y/G), and with a capital output ratio of about 3, and a share of some 10–20 per cent of public capital in total capital, the corresponding rate of return on investment in public capital would be extremely high with an order of magnitude of about 60–130 per cent. This triggered an intensive debate on the plausibility of these estimations, reviewed among others by Gramlich (1994) and Girard et al. (1995).

Issues that were important in the research carried out since then concern among others the type of data used (time series, cross-section, panel data), the way of dealing with dynamics (analysis in terms of levels or first differences), and the spatial level (national versus regional data). Also the possibility that short-term expenditure effects shown in Figure 9.1 are confused with long-run productivity effects played a role in the discussions. See also Sturm et al. (1999) and Romp and De Haan (2007) for a discussion on these issues. Other issues concern the specification of the production function in terms of including factors to take into account business cycle-related issues of capacity utilization, and the use of more flexible functions like the translog function to allow for complementarity between public and private capital (see for example Seitz, 1995).

In the present chapter we focus on some of the more fundamental aspects that have been added, including improving the specification of production function by more explicitly accounting for infrastructure services, causality issues and spatial spillovers.

Specifying infrastructure services

As already indicated, the impact of infrastructure depends on the way it contributes services to economic activities. This calls for a more explicit way of introducing these services (and possibly also its quality into the production function). It will appear that this in general calls for an analysis at lower spatial levels in order to be able to capture the network properties of infrastructure.

An important contribution to the field is provided by Fernald (1999). Instead of using a general indicator of the total value of the capital stock in the production function, he introduces a more refined production function where road transport is explicitly incorporated. The production function for sector i used as a starting point is:

$$Y_i = U_i F^i(K_i, L_i, T(V_i, G)),$$

where U_i denotes the Hicks neutral level of technology, T is an indicator of transport services used in the sector. These services depend on the stock of vehicles in the sector V_i and the capacity G of the road stock provided by the government. Production technology is assumed to be Cobb–Douglas. To account for congestion the road stock capacity indicator is defined as the ratio of total road stock value and vehicle use in terms of total number of miles driven on the roads.

Fernald proceeds by transforming the model into a decomposition of annual productivity changes in terms of changes in the underlying factors of private capital (non-vehicle), labour, the value of the vehicle stock, and the road stock, taking into account the congestion effect. The model implies that sectors with high vehicle intensity would benefit more from increases in the road stock and hence have stronger productivity changes than sectors with low vehicle intensity. Sectoral data for the period 1953–89 in the USA indeed support that vehicle-intensive sectors benefited more than proportionally from road construction programmes, including the emergence of the interstate highway system during the 1950s and 1960s. However, the results also indicate that the contribution of road investments to productivity has decreased in the course of time.

A similar study was carried out by Kopp (2005) for a cross-section of 13 European countries for the years 1976–2000, by focusing on differences between countries; sectoral differences were not considered. A fixed-effects model is used to correct for country-specific unobserved features. He finds a similar positive result for the effect of road investments on productivity changes, but notes that the contribution of the road investments to productivity changes is relatively small.

The step taken by Fernald and Kopp can be considered as a starting point towards a more precise representation of the network features of infrastructure. An obvious limitation of the formulations used is that the transport services involved are modelled in a very imprecise way. Network extension is just modelled by considering the increase in the value of the capital stock. An increase in the road capital stock of a given amount of money X may be the consequence of very diverse projects such as: providing a missing link in a network; linking large cities that were formerly separated; building an urban expressway to counter congestion; or building a connection to a region with low economic potential based on equity considerations. As indicated in section 9.2, the cost of construction per kilometre may vary considerably between regions – in urban areas they tend to be much

larger than in rural ones – and changes in capital stock values must be considered as rather poor indicators of the services provide by infrastructure networks.

Important contributions to address this issue have been made by Forslund and Johansson (1995) and Karlsson and Pettersson (2005) who use the accessibility concept to represent the network properties of transport networks. A typical indicator for accessibility used in this literature would be:

$$Acc_r = \log \Sigma_s \exp \{V_{rs}\} = \log \Sigma_s \exp \{a \cdot X_s - b \cdot c_{rs}\}.$$

This formulation defines the accessibility of region r as the log sum of utilities of interaction with all other regions s. These utilities depend on the relevant qualities of the other regions X_s and on c_{rs}, the interaction costs between r and s, broadly defined. A broad range of accessibility indicators is discussed in Rietveld and Bruinsma (1998). Forms like this allow one to compute the effects of transport network investments in an adequate way. Also congestion effects can in principle be taken on board. Note that this formulation provides a link between the macro-oriented production functions and the modelling of transport networks. It is no surprise, therefore, that accessibility formulations like the one given above also play a role in the field of integrated transport and land use models (see for example Wegener, 2004; Zondag, 2007). Note that the accessibility concept enables the researcher to incorporate the relevant aspects of network morphology.

This approach of using the accessibility concept would in addition bridge the gap, also mentioned by Gramlich (1994), between the general productivity effects generated by the production function literature, representing average conditions within a transport system, and the desire to give advice on the effects of specific infrastructure components. This approach obviously is most convincing when the spatial units are rather small. The reason is that many infrastructure investments have effects that are locally concentrated. When the spatial unit of analysis would be large, most of the effects would then take place within these spatial units, which would mean that they would remain unobserved. However, since databases for spatial research are becoming richer in spatial detail, there are ample opportunities for such an approach.

Spatial spillovers and interdependencies

Various types of spatial spillovers may occur. First of all, infrastructure investments in one region may imply benefits in other regions. The above accessibility indicators related to road infrastructure are an example, but also other types of spillovers may be relevant, such as point infrastructure (for example an airport or a power plant) in one region providing services to neighbouring regions. Other types of spillovers and interdependencies relate to the well-known features of spatial autocorrelation and spatial lags studied in spatial econometric models. A good example of a systematic treatment of these spillovers and interdependencies is given by Kelejian and Robinson (1997). The production function considered here is:

$$Y_{it} = ß_0 \, L_{it}^{ß1} \, K_{it-1}^{ß2} \, G_{it-1}^{ß3} \, D_{it}^{ß4} \, GMN_{it-1}^{ß5} \, PRODMN_{it}^{ß6} \exp[ß7 U_{it} + ß_8 t + ß_{9i} DUM_i + \varepsilon_{it}]$$

Where i and t relate to state and time, D refers to density, GMN is the mean public capital stock in contiguous states, $PRODMN$ is the mean labour productivity in contiguous

states, U is the unemployment rate. Finally, DUM_i is a state dummy for state i to capture unobserved state properties. This model has been estimated with various degrees of econometric sophistication:

0. A reference estimate without fixed effects, that is, without state dummies (ordinary least squares or OLS).
1. 0 + fixed effects (OLS).
2. 1 + temporal autocorrelated disturbance structure in disturbance term ε (nonlinear least squares or NLLS).
3. 2 + explicit recognition of spatial endogeneity of *PRODMN* by means of two stage least squares (2SLS).
4. 3 + explicit recognition of possible endogeneity of L and U by means of 2SLS.
5. 4 + heteroskedasticity of disturbance term ε (generalized method of moments or GMM).

A second series of model estimations has then been carried out to allow for spatial autocorrelation for the disturbance terms.

The 6+6 models have been estimated for US state data for the years 1972–85. Some of the results are rather robust across the various specifications. For example, the order of magnitude of the coefficients for labour β_1 and for the productivity spillover β_6 is rather stable. However, the coefficient for infrastructure β_3 appears to depend strongly on the question of whether or not econometric issues are addressed. For example, in the reference case one gets an estimate of $\beta_3 = 0.15$, but in variants 1–5 negative and significant values are obtained of the order of magnitude of about – 0.15. In the estimations with spatial autocorrelation in the disturbances values are again negative with an order of magnitude of about -0.05, but they are not significant. For the neighbour's infrastructure stock Kelejian and Robinson find mixed results, sometimes negative, sometimes positive, depending on the specification. This leads to the somewhat disturbing conclusion that results of simple econometric methods may be clearly misleading, but that addressing potential spatial econometric problems may well lead to indecisive outcomes.

The spatial econometric approach has found followers in, for example, Lall (2007) who finds significant spillover effects of network investments for the case of India. One of the lessons of the Kelejian and Robinson approach is that improving econometric sophistication is not sufficient to arrive at definitive conclusions. Among the other directions that may be explored is that effort should be devoted to both the specification of the various spillovers – just looking at contiguity is a rather crude way of dealing with network effects – and that the service delivery mechanisms of infrastructure are taken into account in a more explicit way.

Causality issues

One of the issues in the estimation of production functions is that various feedback mechanisms may occur. For example, when considering a time series where a region receives an injection of infrastructure investment, this may indeed have a productivity effect visible in the dependent variable Y, but at the same time it may also lead to the attraction of more labour and private capital, implying that L and K are no longer exogenous. In addition,

the expansion of infrastructure may be the result of a favourable development of GDP in an earlier period, so that the direction of causality is in reality reversed.

Several methods have been used to deal with causality issues. An obvious approach would be to formulate the production function as part of a broader model where public capital formation would be taken on board. The natural starting point would be that public capital formation would be studied from an economic perspective, implying that efficiency considerations play a dominating role. The literature on public choice and political economics provides an alternative perspective suggesting that political considerations like the desire to be re-elected are a main driving force (Persson and Tabellini, 2002).

A nice example of this public choice approach is given by Cadot et al. (2007) who estimate production functions for French regions jointly with investment volumes in the public capital stock using a full information maximum likelihood approach (FIML). They find that public choice-related factors such as the political colour of the region, the congruence between national and regional political colour and the presence of lobby groups indeed play a significant role. Another part of the regional variation in public investments is explained by the development of the nationwide high-speed rail network. An economics-related criterion – rate of return on public investment – has a negative (but limited) effect on the regional volume of public investments. Hence it seems that in France the regional allocation of public investments is mainly driven by political factors and that efficiency considerations play a limited role. A similar result was found for decision-making about road construction projects in Norway (Fridstrom and Elvik, 1997). This combination of economics and public-choice elements in the analysis of infrastructure impacts is apparently an interesting field of analysis. In the case of the Cadot et al. (2007) study, the good thing about the strong role of political factors is that causality issues are of minor concern. Estimation of the key parameters of the production function in the simple one-equation model are hardly different from those in the two-equation model. A similar example of a study on factors influencing the regional allocation of public investments can be found in Lambrinidis et al. (2005) for Greek regions.

An alternative route to explicitly estimating a multi-equation model to deal with causality issues is the use of instrumental variables. This has become a rather standard approach, examples being Holtz-Eakin and Schwartz (1995), Vijverberg et al. (1997) and Percoco (2004). A more complex way to arrive at an integrated analysis of the role of infrastructure in the production process, and possible feedbacks, is the use of a vector autoregression (VAR) approach. In the present context such a model would typically have four endogenous variables: production Y, private capital K, labour L, and infrastructure G. A possible specification for Y would be:

$$Y_t = a_Y + \sum_{j=1}^{p} \left(b_{Y_{t-j}} Y_{t-j} + b_{K_{t-j}} K_{t-j} + b_{L_{t-j}} L_{t-j} + b_{G_{t-j}} G_{t-j} + \varepsilon_{Y_t} \right),$$

implying that the present value of Y is explained by a series of its lagged variables, but also of the lagged values of other endogenous variables K, L and G. Note that the equation above can be interpreted in terms of a production function (for example after applying logs) with the difference that all explanatory variables are lagged. Similar equations are given to explain K_t , L_t and G_t by the same series of lagged variables. This leads to a large number of coefficients to be estimated. In particular, note that when p, the number

of lags increases with 1, the additional number of coefficients to be estimated equals 4 for the above equation, and 4x4 for the whole system of equations. As explained by Greene (2003), a possible perspective on VAR models is to interpret them as the reduced form of complex simultaneous equation models. An obvious advantage of VAR models is that there is no need to decide on which of several contemporaneous variables is endogenous in a model. VARs can be used for forecasting, testing of Granger causality, and simulations of policy alternatives, such as applying a certain shock in infrastructure.

Two important themes are relevant in this context. The first is to what extent there is still a significant effect of infrastructure on production, and what is the order of magnitude. The second is to what extent we observe reverse causality issues like developments in production value affecting levels of infrastructure capital formation. For example Pereira and Flores de Frutos (1999) find that infrastructure has a significant impact on production, but much smaller than indicated by the early results of Aschauer. Another example of a study in this field is Pereira and Andraz (2006), who find for Portugal that the effects of public investments depend strongly on the type of region, and in particular that the capital region of Lisbon benefits more than proportionally from public investments compared with the other regions in the country.

A broader survey on the use of VAR approaches to infrastructure effects is given by Romp and De Haan (2007). They find that most studies conclude that there is a positive long-run effect of infrastructure shocks on production levels; and also that many studies find examples of significant reverse-causality effects of production on public capital formation. Thus, the VAR approach can be considered as a very useful addition to the original single-equation studies, leading to added insights about the value of the output elasticity of infrastructure while correcting for possible reverse-causality problems.

Crowding-out

Crowding-out means that an increase in government spending leading to an increase in interest rates will discourage private investment. This effect, which is clear according to most theoretical approaches, has been tested by Easterly and Robelo (1993) for a long time series for some 30 countries, but the results are rather fragile. A more recent contribution is that of Ahmed and Miller (2000), who find that, based on panel data for some 40 countries, government expenditure items in general crowd out investment, with the exception of transportation and communication expenditure, which crowds in investment, especially in developing countries. In developed countries, no significant effect is found. This difference between developing and developed countries may indicate that the strongest productivity-enhancing effects of transport infrastructure investments occur during the earlier phases of economic development, whereas after that such effects are more moderate. This is confirmed by Fernald (1999) who finds for the USA that road construction in the 1950s and 1960s boosted productivity, but effects at later stages were much smaller.

Another point that deserves attention here is that the degree of crowding-out will depend on the way in which the expenditures are financed. For example, Ahmed and Miller (2000) find differences between debt-financed and tax-financed expenditures, tax-financed expenditures having larger crowding-out effects. A similar result is found by Kim (1998) who uses a CGE model to analyse the macroeconomic effects of transport infrastructure investment.

9.5 Optimal provision of infrastructure

After reviewing the literature on the productivity effects of infrastructure supply we now turn to a review of the welfare effects. Let us, as a point of reference, review the fundamentals of optimal provision of infrastructure by addressing rules of allocating road capacities efficiently. These rules were first stated in the 1960s by Mohring and Harwitz (1962) and Vickrey (1969) and have been generalized in Buchanan's (1965) theory of clubs (see Sandler and Tschirhart, 1997, for the theory of clubs). For more extensive reviews see Small (1992) or Arnott and Kraus (2003). A road is a collective-use good, neither a pure private nor a pure public good. The collective character is due to the fact that rivalry of users is limited. Rivalry can be absent, for example on a highway through rural areas or at night. It increases with increasing congestion, which is itself increasing in the intensity of use and decreasing in the capacity installed. Optimal allocation has two sides: the capacity to be installed, and the intensity of use. For an optimal allocation, both aspects have to be optimized. As the decision to use the capacity is decentralized, price incentives are needed for attaining optimal use.

For the sake of simplicity, consider a single road and a set of identical users, each enjoying benefit $B(x)$ if using the road with intensity x (number of trips per user). The benefit is measured in monetary terms; it is the willingness to pay for being able to use the road with intensity x. Demand is distributed uniformly in time. The user faces congestion cost $c(nx, K)$ per trip. c is also measured in monetary units, that is time is translated to money. n is the number of users, K is the cost of capacity.[1] c is increasing in nx and decreasing in K. The total net welfare of the typical user is thus:

$$W^0(x, K, n) := B(x) - xc(nx, K) - K/n.$$

The setting resembles the club model of Berglas (1976). Maximizing simultaneously with respect to all three arguments gives the first-order conditions:

$$B_x = c + nxc_1, \tag{9.1}$$

$$-nxc_2 = 1, \tag{9.2}$$

$$x^2nc_1 = K/n. \tag{9.3}$$

(9.1) is the condition of optimal use, saying that intensity x has to be chosen such that the marginal benefit B_x (the derivative of B with respect to x) has to be equal to the marginal social cost, which is average private cost c plus congestion externality of a trip nxc_1. c_1 and c_2 denote the partial derivatives of c with respect to the first and second argument of the congestion cost function. Equation (9.2) is the optimal investment rule stating that capacity should be extended to the point where the marginal cost of capacity (which, by the definition of capacity, equals unity) equals the joint marginal congestion cost saved by all users. The final condition (9.3) is the optimal membership condition. It states that the average cost per 'club member', K/n, equals the total marginal congestion cost induced by an extra club member. Conditions (9.1) to (9.3) jointly determine what Buchanan has called an optimal club; optimality is seen from the 'within-club viewpoint'. It is also optimal for the society as a whole if the entire population N partitions into groups of size n^* without a remainder, n^* denoting the optimal number of members. N/n^* is then the number of clubs, that is, the number of parallel links in case of roads. It is however

unlikely that the technology renders more than one parallel road optimal (see below). If clubs of size n^* leave a remainder, the within-club and societal viewpoints differ, as the society as a whole also cares about those left non-served by optimal clubs. For the entire society a natural objective is treating all individuals alike, as they are all identical. A social planner would thus either increase the club size or decrease it and add one club, such that everyone is served. Which alternative to choose depends on which offers a higher benefit.

As we will show now, under the given assumptions an optimal road is self-financing, if users pay a fee per trip just covering the externality, that is, nxc_1. Leaving the decision about x to the user, she chooses x such that B_x equals private unit cost c plus fee per trip, which means that (9.1) holds. Interestingly, the revenue from the fee paid by all users, $(nx)^2c_1$, just covers the optimal capacity cost, which is immediate from multiplying (9.3) by n. This demonstrates the self-financing property. Furthermore, optimality with respect to both capacity and pricing implies local zero-homogeneity of c. Local zero-homogeneity of c at a given level of K and nx means that a 1 per cent increase of use nx and a simultaneous 1 per cent increase of K leaves congestion unchanged. To see that local zero-homogeneity holds, use $(nx)^2c_1 = K$ by (9.3) and $nxKc_2 = -K$ by (9.2) to get $(nx)^2c_1 + nxKc_2 = 0$. As $nx \neq 0$, this is equivalent to $nxc_1 + Kc_2 = 0$, which is the formal condition of local zero-homogeneity of c. If c is globally zero-homogenous (that is, zero-homogeneous at any point), then conditions (9.1) and (9.2) define optimal use and optimal capacity, respectively, and (9.3) holds for any choice of n; one can build many narrow roads or one wide road without affecting the benefit. The optimal road price always just covers the capacity cost.

We now discuss the implications of the above findings for private versus public provision of infrastructure. Global zero-homogeneity would be an attractive property of c, as it implies that we only would have to find the optimal fee (requiring of course to solve equations 9.1 and 9.2 for both, the optimal x and K), and then could advise the road administration just to exhaust the budget. Even better, if either zero-homogeneity holds globally or the optimal club size n^* exists and is small relative to the population total N, and if furthermore users not paying a fee are excludable, then provision of optimal roads can be left to the market. As long as the population is not partitioned into optimal clubs, there is room for providing a new road with fees making users better off and leaving a profit to the provider. Providers enter the market until the optimal structure is attained and profits are competed away.

At first sight, this is not applicable to roads anyway due to the technical problems of raising a fee and excluding non-payers, at least for roads other than motorways. But things will change rapidly. In a couple of years any new car will have sufficient intelligence on board to make even highly complicated pricing schemes, possibly varying by road, time of day, actual congestion, and so on, technically feasible. Though this may in fact lead to more club-like private provision of roads, first-best optimality in the above sense will not be achieved.

The first and main difficulty – assuming cheap excludability to be possible – is that the optimal club size will in reality often be large relative to N, possibly bigger than N or even infinite. The latter holds under global economies of scale. Let $W^a(n)$ denote the maximal welfare attainable for the typical user, if the club size is fixed at n:

$$W^a(n) := \max_{x,K} W^0(x, K, n).$$

The optimal club size is n^* maximizing $W^a(n)$ subject to $n \leq N$. If W^a is increasing for all $n \leq N$, the optimal club size is N, and $W_n^a(N) > 0$. $W_n^a(N)$ denotes the derivative evaluated at N. Invoking the envelope theorem and taking derivatives of W^0 with respect to n at x^*, K^*, N, where x^* and K^* are optimal for $n = N$, yields:

$$W_n^a(N) = -x^{*2}c_1 + K^*/N^2.$$

Thus $W_n^a(N) > 0$ implies $K^* > (Nx^*)^2 c_1$ so that optimal fees do not cover the cost. Also $Nx^*c_1 + K^*c_2 < 0$, saying that the local degree of homogeneity is negative, saying in turn that congestion costs per trip decline if both total use and provision cost simultaneously increase by the same percentage. It is unlikely that n^* is less or even much less than N. Otherwise we should often observe the building of parallel roads. That we rarely see them can hardly be explained by inefficient provision; the only sensible explanation is economies of scale.

Thus private road providers turn out to be natural monopolists, and the theory of natural monopoly (Tirole, 1988) can be applied in the discussion of possible institutional settings allowing good (though not first-best) allocation. Unfortunately, the hope that, despite the natural monopoly, private providers could be forced to efficient provision by market contestability is in vain. The assumption of zero sunk costs required for contestability is obviously vastly at odds with the facts in road transport.

Given however that natural monopolies do exist in other industries with non-zero sunk costs, and that they are nevertheless disciplined by a mixture of providers of substitutes, threat of entry and possibly also some kind of regulation, one can well expect private provision to expand in the future, once exclusion costs have declined to close to nothing. And the efficiency of market provision, though far from perfect, may well outperform that of public provision, suffering from administrative failure of all kinds. Private provision is of course facing a lot of additional problems, beyond natural monopoly, such as lack of information, non-acceptance of complex pricing schemes and so forth. But none of those is specific to the private provider; public administration is facing the same problems. Therefore the general conclusion of increasing opportunities for private provision under low exclusion cost remains true.

Still, public provision will continue to dominate for some decades. Therefore we analyse the optimality conditions for public supply in more detail. A public provider should try to maximize $W^0(x, K, N)$ with respect to x and K. This requires a road price Nx^*c_1 per trip and provision according to the rule $-Nx^*c_2 = 1$, where x^* is the intensity chosen under an optimal road price. The objective function is thus:

$$W^b(K) = \max_x W^0(x, K, N).$$

Under $W_n^a(N) > 0$, as assumed before, maximizing W^b generates the deficit $K - (Nx^*)^2 c_1 > 0$, which must be financed by general taxes. If distortions of the tax system are taken into account, the optimal provision rule is modified to $-Nx^*c_2 = \lambda$, where $\lambda > 1$ denotes marginal cost of public funds.

We have just argued that efficient user charges cannot be applied for technical and other reasons, at least not yet everywhere. Hence, public provision is unable to achieve the maximum of W^b. Without user charges it has to take $W^c(K)$ as an objective, defined as:

$$W^c(K) := W^0(\tilde{x}, K, N),$$

with $\tilde{x}$ denoting the non-optimal decisions users make when facing only private instead of social marginal congestion cost, that is, when $\tilde{x}$ is such that $B_x = c$.

Assume the provider has a model at hand allowing to estimate c as well as users' demand responses correctly. How could he identify W^b or W^c without directly observing benefits B? Dupuit's ingenious answer is the surplus function $S(p)$, defined as:

$$S(p) = \max_x \{B(x) - px\}$$

with observed generalized cost p per trip. As $S_p(p) = -x(p)$ is observed demand, changes of S can be inferred on from changes of p by integrating $-x(p)$. Using S, W^b is rewritten as:

$$W^b(K) = S(p) + ex - K/N$$

with $p = c + e$, externality charge $e = Nxc_1$, and $x = -S_p$. Similarly we get:

$$W^c(K) = S(c) - K/N.$$

Both objectives easily carry over to the realistic heterogeneous case where users are allowed to be all different. Define $W^0(x, K)$ as the average net benefit per user (the argument N is now hidden, because it is not variable):

$$W^0(X, K) := \frac{1}{N}\sum_i \left[B_i(x_i) - x_i c_i(x., K)\right] - K/N.$$

$X = (x_1, \ldots, x_N)$ is the vector of individual intensities, c_i is the congestion individual i is facing; it depends on K and $x. := \sum_i x_i$.

$$W^b(K) := \max_X W^0(X,K)$$
$$= \overline{S}(p) + e\overline{x} - K/N.$$

with averages

$$\overline{S}(p) := \frac{1}{N}\sum_i S_i(p_i) \text{ and } \overline{x} := x./N,$$

and with the generalized cost $p_i = c_i + e$ and externality $e = \sum_i x_i c_{i1}$. c_{i1} is the derivative of c_i with respect to its first argument. Note that e does not vary across users, though c_i does. If congestion functions differ across users due to different values of time, knowledge of total use $x.$ is not sufficient for identifying W^b, because the whole vector matters for e. In practice one has to rely on approximations such as $e \approx x.\overline{c_{i1}}$. W^c is simply:

$$W^c(K) = \overline{S}(c) - K/N$$

with obvious notation. Thus, optimal pricing and capacity rules can be derived with heterogeneous agents as well, and self-financing can be shown still to hold under constant returns to scale.

How about a network of roads? In case of efficient use and small projects a road can be evaluated separately without caring about effects in the rest of the network (Arnott and Kraus, 2003, p. 711). This is an implication of the envelope theorem as we will show now. Let y be a vector of intensities on roads other than the one to be evaluated, and define W^0 now in an obvious way as:

$$W^0(x,y,K) := \frac{1}{N}\sum_i \left[B_i(x_i,y_i) - x_i c_i^x(x.,K) - y_i c_i^y(y.)\right] - K/N.$$

c^x and c^y denote congestion cost on the studied road and on other roads, respectively. K is the cost of capacity of the studied road. The correct objective function for the public provider is now $W^b(K)$, defined as:

$$W^b(K) := \max_{x,y} W^0(x,y,K),$$

while in practice one would likely use a criterion $\tilde{W}^b(K)$, that coincides with $W^b(K)$ at a reference situation K^0, for which x^0 and y^0 are optimal, while for $K \neq K^0$, x and y do not fully adjust to the optimum. For example, one neglects the response of y by fixing it at y^0, or one calculates responses of x and y holding fees constant at their reference values. Formally speaking, we know that $\tilde{W}^b(K) \leq W^b(K)$ with equality if $K = K^0$. Hence:

$$\tilde{W}^b_K(K^0) = W^b_K(K^0) = -\left[\sum x_i^0 c_{i2}^x(x.^0,K^0) + 1\right]/N$$

while $W^b(K) - W^b(K^0) \geq \tilde{W}^b(K) - \tilde{W}^b(K^0)$. A small project is evaluated correctly by just comparing congestion cost savings on the road affected by the project, with project cost, keeping reference intensity constant, while welfare gains of big projects are underestimated.

The fact that one focuses on just one road under efficiency of the reference situation looks like a great simplification, but it is only superficially so, because the reference situation itself can only be found if optimal user fees are known; these in turn can only be found by optimizing the whole network.

Apart from the fact that it would be a demanding task to determine the optimal capacities and congestion charges for a whole network, we have argued before that efficiency of the reference situation is unlikely, because it is an exception that congestion externalities are priced. This means that public providers usually follow objective W^c rather than W^b. An isolated evaluation of a project by W^c gives incorrect results even for small projects, because congestion externalities in the entire network must not be neglected.

Extending W^c to a network is conceptually largely a matter of notation, while empirical implementation is another matter. Let x_{ij} denote trips of user i along road j, $x_i := (x_{i1}, \ldots, x_{il})$ the vector of trips along roads 1, …, l, taken by user i, and $x.: = \Sigma_i x_i$, the vector of the total number of trips along roads 1, …, l. $K := K_1, \ldots, K_l$ is the vector of costs for providing road capacities on roads 1, …, l. Finally $c_i := c_{i1}, \ldots, c_{il}$ is the vector of generalized costs of user i for trips along roads 1, …, l. Then:

$$W^c(K) := \frac{1}{N}\left[\sum_i S_i(c_i) - \sum_j K_j\right]$$

with

$$S_i(c_i) = \max_{x_i} \{B_i(x_i) - x_i \cdot c_i\},$$

$$c_i = c_i(x., K) \text{ and } x_{ij} = -\frac{\partial S_i}{\partial c_{ij}}.$$

Finding an equilibrium means to solve for x and c simultaneously, where x depends on c and c depends on x (and on K). Specification of an equilibrium can start with specifying the benefit. This is the case in a stochastic user equilibrium model, where demand is derived from stochastic benefit maximization. If instead one starts from specifying demand $x_i(c_i)$, one has to obtain a surplus change $S_i(c_i^1) - S_i(c_i^0)$ from integration:

$$S_i(c_i^1) - S_i(c_i^0) = \int_{c_i^1}^{c_i^0} x_i(c_i) \cdot dc_i.$$

Note that now x_i, c_i, dc_i, c_i^0 and c_i^1 are vectors of length l, respectively, and the dot denotes the inner product. The integral is understood as a line integral. A well-defined surplus exists only if this line integral is independent of the integration path. Furthermore, the definition of the surplus implies that it is a convex function, because it majorizes its tangential planes. This in turn implies that the function $x_i(c_i)$, if differentiable, must have a symmetric negative-semidefinite Jacobian. Otherwise, using line integrals as surplus measures lacks a theoretical underpinning (Mas-Colell et al., 1995, Chapters 3.H and 3.I). For a small project one has simply $dS_i = -x_i \cdot dc_i$; net gains from demand response can be neglected. Assessing a road project in isolation does not give correct answers, even for small projects, because investing in road j affects congestion on other roads not neutralized by externality fees. As a rule, the sum of isolated assessments of road segments overestimates or underestimates a joint assessment, depending on whether the segments are serial or parallel.

9.6 CGE analysis of regional effects of infrastructure investments

Thus far, we have discussed how to quantify the benefits from using transport infrastructure without caring about who would be the final beneficiary, and in particular where they are located. Though we can measure the benefit by calculating the surplus of the direct user of the infrastructure component, this does not mean that this user is the final beneficiary. How could one identify where the benefits eventually go?

The most advanced methodology currently available is to set up a spatial computable general equilibrium (CGE) model, in which interregional flows of goods and passengers are explicitly modelled, and in which the equilibrium can be shocked by varying the transport costs. At the same time, these models can also identify the regional welfare impact generated by constructing and financing the infrastructure. Finally, impacts of pricing schemes may also be studied (see Bröcker, 2004, for a review of CGE models in transport).

A computable general equilibrium model is a textbook general equilibrium model 'filled with numbers'. Filling it with numbers means that general concepts of the theory such as utility and production functions are replaced with specific parametric functional forms and concrete numbers are assigned to the parameters. Typically, the functions contain two types of parameters: position parameters shifting supply and demand schedules to the left or right; and elasticities determining the slopes, that is, direct as well as cross-price responses of supply and demand. Position parameters are calibrated, which

means to fix them such that social accounting data of a reference year such as output, factor incomes, consumption, investment, tax revenues and public expenditure and, most importantly, interregional trade are reproduced by the equilibrium solution representing that reference situation. Elasticities cannot however be calibrated from a one-shot accounting database and thus have to be imported from econometric studies. As these studies rarely refer to exactly the same industries, commodities and economic environments as the ones under study, determining elasticities by 'literature search' in this way is clearly a weak point of the approach.

CGE models can have very different degrees of complexity in terms of number of industries, representative consumers, regions and points in time admitted, as well as in terms of market structures taken into consideration. The minimum requirements that all CGE models have in common are:

- Income–expenditure consistency, meaning two things: first, each agent's (firm, household, state) revenue must equal their expenditure; and second, all expenditure (purchases, taxes, and so on) of an agent reappear as other agents' revenues (sales, tax revenues, and so on).
- Rationality, meaning that behaviour of any agent is derived from an explicit optimization approach (possibly with the exception of the state(s)).
- Equilibrium, meaning that the economy is supposed to rest at a point where no agent has reason to revise their decision, and decisions are mutually consistent.

Models with many regions appeared early in the development of the approach in an international trade context (Shoven and Whalley, 1984). First-generation models were static with perfect competition; imperfect markets, time, forward-looking agents and random shocks have been introduced in later developments (Ginsburgh and Keyzer, 1997).

An obvious strength of this approach is that it directly lends itself to welfare analysis of policy measures, because representative households explicitly aim at maximizing utility, which gives a natural welfare criterion. Utility itself is an ordinal concept and thus not usable directly for evaluation in monetary terms, but utility changes can be translated into monetary amounts by the Hicksian concepts of compensating or equivalent variation. The latter is most often used, measuring the amount of money one would have to transfer to a person in the reference situation in order to make them as well off as they are going to be in the alternative. It is important to understand that in this approach only private households are the ones to whom net benefits eventually accrue. Any producer surplus that appears on the firm's side in standard partial analysis is eventually transferred to private households, either through changes in factor incomes, or through price effects, or through profit transfers to shareholders.

In order to be useful for identifying net benefits of changes in transport infrastructure networks by region, a CGE must be of the multi-regional type. For identifying benefits due to freight cost reductions, it must explicitly model interregional trade, and for identifying benefits of passenger travel cost reductions, it must explicitly model interregional passenger travel flows. Costs must contain monetary as well as time components, which in the case of private households means that the decision model must take the monetary as well the time budget of the decision-maker into account.

Models available in the literature only partly fulfil these broad requirements. Models covering long-distance trade are well developed, but they lack private passenger travel (Bröcker, 2002; Knaap and Oosterhaven, 2002, 2004). Other models focus on passenger travel on an urban scale (Anas and Kim, 1996). Models covering both, and adequately representing the interregional as well as the urban dimension, have still to be developed. Furthermore, we still do not yet see a full integration of equilibrium flows through a congested network into CGE models. Thus, there is a certain division of labour between transport engineers on the one hand assessing projects by surplus measures in conventional four-level transport network models, and CGE modellers on the other hand receiving cost information from the transport engineers and feeding them into CGEs that allow focus on the spatial distribution of benefits as well as 'wider economic effects'. The latter are additional (positive or negative) effects in other parts of the economy, that are not covered by the conventional surplus measure in case of imperfect markets (see section 9.8 below).

9.7 A gravity approach to regional welfare measurement of infrastructure investments

Obviously, setting up CGE models that are both realistic as well as founded in microeconomic theory is still a formidable task. A reasonably good first approximation can however be obtained with considerably less effort in terms of data and computational complexity by a partial equilibrium approach. The theory of partial spatial price equilibrium began with Samuelson (1952) and was given a full account of by Takayama and Judge (1971). Here we propose a recipe confined to identifying regional welfare effects of road use for freight. A passenger model can be set up in similar spirit. The idea is to take into account only the effects in regions of origin and destination of interregional flows. This is a short-cut because some of the benefits assigned to the regions this way are eventually transmitted to other regions by input–output, income and final demand linkages. This 'second round' as well as any further round of benefit redistribution is neglected. We focus on functional forms that lead to gravity-type specifications of equilibria, which have been shown to perform extremely well empirically. Traditional gravity specifications are surveyed by Fotheringham and O'Kelly (1989). The relation between gravity flows and spatial price equilibrium was for the first time discovered by Golob and Beckmann (1971), and rediscovered at least once in international trade theory (Anderson, 1979; for a more recent treatment see Anderson and van Wincoop, 2003).

To be concrete, take the case of goods transport through a network with known costs per unit equal to c_{rs} for flow quantity x_{rs} from region r to region s. Let p_r denote the price per unit at the origin and $q_{rs} := p_r + c_{rs}$ the price per unit of a good from r at destination s, including transport cost (inclusive price). For a destination s we collect prices by origins $1,\ldots,m$ in a vector $q_s := (q_{1s},\ldots,q_{ms})$. Let agents in the origins and destinations be price takers with supply functions $S_r(p_r)$ and demand functions $D_s(q_s) := D_{1s}(q_s), \ldots, D_{ms}(q_s)$, respectively. S_r is $\mathbb{R}\to\mathbb{R}$, assigning a scalar supply quantity to a scalar price, while D_s is $\mathbb{R}^m\to\mathbb{R}^m$, assigning the vector of demand by origin to the vector of inclusive prices by origin. Note that each component function depends on the entire vector.

The tools for assessment are the surplus functions, associated with each origin and destination. The supply surplus is a function $P_r(p_r)$, unique up to an additive constant, with $\mathrm{d}P_r(p_r)/\mathrm{d}p_r = S_r(p_r)$, that is, the integral over supply. It is convex because S is supposed to be non-decreasing. The demand surplus is a convex $\mathbb{R}^m\to\mathbb{R}$ function $C_s(q_s)$ with

$\partial C_s(q_s)/\partial q_{rs} = -D_{rs}(q_s)$. As already mentioned in section 9.5, demand must have a symmetric negative-semidefinite Jacobian in order that this surplus be well defined and microfounded.

The partial spatial price equilibrium is attained if supply in each origin equals demand for goods from the respective origin, that is, $S_r(p_r) = \sum_s D_{rs}(q_s)$ for all r. Now consider an exogenous change of transport cost from c_{rs}^0 in a reference situation to c_{rs}^1 in an alternative situation. The task is to evaluate the welfare impact of this change by region. The method is straightforward: solve the equilibrium equations for both the reference and the alternative, and calculate the surplus changes $P_r(p_r^1) - P_r(p_r^0)$ and $C_s(q_s^1) - C_s(q_s^0)$.

Convenient functional forms for implementing this approach are the logit-exponential and the CES-power forms. The interesting conclusion to be drawn is that with these specifications we can infer welfare effects by region from reduced-form solutions that do not contain prices any more. This is a big advantage, because price information is hard to obtain. The logit-exponential is the form of choice in case we have quantity observations of flows such as tonnes, the CES-power is the form of choice in case of value information. Anyway, both are close relatives as shown by Anderson et al. (1988). We start with the logit form (Domencich and McFadden, 1975), assuming that supply is an exponential in price. Similarly demand, aggregated over origins, is an exponential in a composite price, with an appropriate definition of composition, while the split of the aggregate across origins is controlled by a logit. Formally, $S_r(p_r) = a_r \exp(\alpha p_r)$, $\overline{D}_s(\bar{q}_s) = d_s \exp(-\gamma \bar{q}_s)$, and:

$$D_{rs}(q_s) = \frac{b_r \exp(-\beta q_{rs})}{\sum_t b_t \exp(-\beta q_{ts})} \overline{D}_s(\bar{q}_s),$$

with

$$\bar{q}_s = -\frac{1}{\beta} \log \sum_r b_r \exp(-\beta q_{rs}). \tag{9.4}$$

a, d and b are position parameters later dropping out of the reduced-form solution. α, β and γ are semi-elasticities, measuring the relative change of the quantities per absolute change of price or unit cost. Note that they are not dimensionless. Their dimension is the inverse of that of unit costs or prices; if the latter are in euros per tonne, say, then the semi-elasticities have dimension tonnes per euro.

The corresponding surplus functions are:

$$P_r(p_r) = \frac{a_r}{\alpha} \exp(\alpha p_r)$$

and

$$C_s(q_s) = \frac{d_s}{\gamma} \exp(-\gamma \bar{q}_s).$$

For checking the latter formula, just take the derivative with respect to q_{rs}, using the chain rule and equation (9.4). Convexity can also be shown.

Equilibrium flows are $x_{rs} = D_{rs}(q_s)$ with $\Sigma_s x_{rs} = S_r(p_r)$. If we know origin and destination totals $x_{r.}^0 := \Sigma_s x_{rs}^0$ and $x_{.s}^0 := \Sigma_r x_{rs}^0$ for the reference situation, the reference flows solve the doubly constrained gravity model:

$$x_{rs}^0 = A_r B_s \exp(-\beta c_{rs}^0),$$
$$\sum_s x_{rs}^0 = x_{r.}^0$$
$$\sum_r x_{rs}^0 = x_{s}^0.$$

The multipliers A_r and B_s gather all variables with subscript r and s, respectively, and are determined by the two constraints. Hence, knowledge of the parameter β, generalized transport costs c_{rs}^0 and the origin and destination totals suffices for solving for the reference flows. Knowledge on prices is not needed: they are incorporated in the multipliers A_r and B_s.

The β-parameter is easily estimated by representing A_r and B_s as fixed effects. An obvious idea is to take logs and apply ordinary least squares (OLS), but non-linear estimation of the untransformed model is to be preferred, as it allows for zeros in the data and better accounts for heteroskedasticity, that is usually observed in the residuals (Bröcker and Rohweder, 1990; their approach has recently been reinvented by João et al., 2006). Once reference flows are known, the alternative equilibrium resulting from a change in transport costs can be rewritten as:

$$x_{rs}^1 = x_{rs}^0\left[(\beta - \gamma)\Delta\bar{q}_s - \beta(\Delta p_r + \Delta c_{rs}\right],$$
$$\sum_s x_{rs}^1 = x_{r.}^0 \exp(\alpha \Delta p_r),$$
$$\sum_r x_{rs}^1 = x_{.s}^0 \exp(-\gamma \Delta \bar{q}_r),$$

with $\Delta c_{rs} := c_{rs}^1 - c_{rs}^0$ and so forth. This system of equations allows us to compute the price changes resulting from a change in transport costs. For solving it one has to know the reference flows, the cost changes and two additional parameters, the semi-elasticities α and γ. Their estimation turns out to be more difficult. The model can be shown to be structurally identical with Alonso's 'theory of movement' (Alonso, 1978). Therefore an instrumental variable estimator of Alonso's model proposed by de Vries et al. (2002) can be applied, that works without price information. Another solution is to rely on supply and demand price elasticity estimates from the literature, which are translated to semi-elasticities by dividing through average unit prices. Note that elasticities are dimensionless, such that dividing them by the price (euros per tonne, say), renders a parameter in tonnes per euro, as required.

The price changes in turn uniquely determine the surplus changes:

$$P_r(p_r^1) - P_r(p_r^0) = \frac{x_{r.}^0}{\alpha}[\exp(\alpha \Delta p_r) - 1],$$

and

$$C_s(q_s^1) - C_s(q_s^0) = \frac{x_{.s}^0}{\gamma}[\exp(-\gamma \Delta \bar{q}_s) - 1].$$

The supply surplus gain is monotone increasing in the price increase, the demand surplus gain is monotone decreasing in the price increase. One can rewrite these indicators in a

form showing that surpluses are reference flows multiplied by a transformation of the percentage change of a demand potential and a supply potential, respectively. The bigger the potential increase, the larger the welfare gain.

We will show now that a similar approach can be followed with the CES form. In the CES-power specification we assume a unit of good used in destination s to be a CES composite of goods distinguished by place of origin. Customers choose the composition that minimizes expenditure per unit of composite, given inclusive prices, gathered in vector q_s defined as above. Supply and demand are now power functions: $S_r(p_r) = a_r p_r^{\alpha}$ is supply, $D_s(\overline{q}_s) = d_s \overline{q}_s^{-\gamma}$ is demand for the composite good as a function of the price for the composite good. This price is the minimal expenditure per unit of the composite, which is:

$$\overline{q}_s = (\sum_r b_r q_{rs}^{1-\beta})^{\frac{1}{1-\beta}}. \tag{9.5}$$

As before, a, d and b are position parameters later dropping out of the solution. α, β and γ now are dimensionless elasticities measuring percentage quantity changes per percentage price change. While quantity shares are proportional to exponentials of inclusive prices in the logit, value shares are proportional to power functions of inclusive prices in the CES:

$$q_{rs} D_{rs}(q_s) = \frac{b_r q_{rs}^{1-\beta}}{\sum_t b_t q_{ts}^{1-\beta}} \overline{q}_s \overline{D}_s(\overline{q}_s).$$

Using (9.5) this can be simplified to:

$$D_{rs}(q_s) = b_r \left(\frac{q_{rs}}{\overline{q}_s} \right)^{-\beta} \overline{D}_s(\overline{q}_s).$$

Note that adding up values from all origins to a destination s yields the value of the composite in s, while adding up the quantities does not make sense, because goods from different origins are supposed to be different. To the contrary, in the logit quantities add up, but values do not.

The corresponding surplus functions are:

$$P_r(p_r) = \frac{a_r}{1+\alpha} p_r^{1+\alpha}$$

and

$$C_s(q_s) = \frac{d_s}{\gamma - 1} \overline{q}_s^{1-\gamma}.$$

As above, the proof is to take the derivative with respect to q_{rs}, using the chain rule and equation (9.5).

We now introduce the famous iceberg assumption, stating that transport costs come in the form of 'melting' of goods on their way from origin to destination by a factor $t_{rs} > 1$. If one unit is sent off from r, only $1/t_{rs}$ units arrive, and if the price per unit at origin r is p_r, then the price at the destination s is $q_{rs} = p_r t_{rs}$. This somewhat strange assumption provokes critical objections (McCann, 2005). It can be demonstrated that the approach also goes through without it, but at the cost of making the equations less elegant and making the

close relation to the logit less lucid. A nice implication of the iceberg assumption is that we need not care whether values are meant to be in prices of the origin or the destination: they are both the same.

Let $y_{rs} = p_r t_{rs} D_{rs}(q_s) = q_{rs} D_{rs}(q_s)$ denote the value of flows from r to s, and let $y^0_{r.} := \sum_s y^0_{rs}$ and $y^0_{.s} := \sum_r y^0_{rs}$ be the marginal totals of reference values. Again gathering terms with subscript r and s in A_r and B_s, respectively, the reference value flows also fulfil a doubly constrained gravity model:

$$y^0_{rs} = A_r B_s (t^0_{rs})^{1-\beta},$$

$$\sum_s y^0_{rs} = y^0_{r.},$$

$$\sum_r y^0_{rs} = y^0_{.s}.$$

Furthermore, let ratios of alternative over benchmark prices, costs, and so on be marked by hats, $\hat{p}_r := p^1_r / p^0_r$ and so forth. Then everything is written in terms of benchmark values and ratios as follows:

$$y^1_{rs} = y^0_{rs} (\hat{\bar{q}})^{\beta-\gamma} (\hat{p}_r \hat{t}_{rs})^{1-\beta},$$

$$\sum_s y^1_{rs} = y^0_{r.} (\hat{p})^{1+\alpha},$$

$$\sum_r y^1_{rs} = y^0_{.s} (\hat{\bar{q}})_{1-\gamma}.$$

The similarity to the logit-exponential system is obvious. This can be solved for the relative price changes, which in turn uniquely determine the surplus changes:

$$P_r(p^1_r) - P_r(p^0_r) = \frac{y^0_{r.}}{1+\alpha}\left[(\hat{\bar{p}}_r)^{1+\alpha} - 1\right],$$

and

$$C_s(q^1_s) - C_s(q^0_s) = \frac{y^0_{.s}}{\gamma - 1}\left[(\hat{\bar{q}}_s)^{1-\gamma} - 1\right].$$

As before, the supply surplus gain is monotone increasing in the price increase, the demand surplus gain is monotone decreasing in the price increase. One can also rewrite these indicators in a form showing that surpluses are reference flows multiplied by a transformation of the percentage change of a demand potential and a supply potential, respectively. The bigger the potential increase, the larger the welfare gain.

These derivations for the logit and CES forms demonstrate that the gravity approach, which is widely used in the domains of international trade, regional economics and transport economics, can be used to study the welfare effects of changes in transport costs. Hence, when this is the aim of the analysis, it provides an attractive alternative to the more involving construction of CGE models.

9.8 Surplus equivalence in welfare analysis

So far we have assumed changes in the sum of surpluses, measured either directly on the road network or on the regional level, to quantify correctly changes in social welfare generated by project use. It is now well understood that this in fact holds true, if allocation in the economy is efficient (Lakshmanan et al., 2001, section 2.4). This fact is called 'surplus equivalence' (SPE) in transport economics. Economists sometimes call this fundamental insight the 'the cost of a cost is its cost' theorem. If at the margin the cost saving in the transport network is $-\Sigma_i x_i^0 \cdot dc_i$ for fixed reference intensities x_i^0, then this is exactly what the society as a whole saves. Though this is often misunderstood, it is close to trivial: if allocation is efficient, one can clearly not save less, because this is what one saves even without any adjustment of allocation. But given a marginal change one also cannot gain more, because in an efficient allocation one cannot gain anything by marginal reallocation; otherwise the allocation was not efficient. There are many reasons why the conditions for SPE do not hold in a real economy. Without claiming completeness, we just enumerate a few reasons why a transport cost reduction can generate extra gains or losses not covered by the traditional surplus measure:

1. Sectoral shift: different industries are characterized by different deviations of marginal willingness to pay (MWTP) from marginal social costs (MSC) due to different degrees of market power. Output expansion (contraction) in industries with excess of MWTP over MSC causes extra gains (losses) not covered by the surplus measures (Venables and Gasiorek, 1999).
2. Employment shift: under unemployment there is a gap between marginal social return of labour and its opportunity cost. Hence, a shift of labour demand from low unemployment to high unemployment places generates extra gains, and vice versa.
3. Lower trade costs enforce competition, thus bringing MWTP and MSC closer to one another, thus generating extra welfare gains. While the net effects according to reasons 1 and 2 are ambiguous, the extra gain of a transport cost reduction is in this case always positive.
4. Under oligopolistic conditions there is a tendency to wasteful reciprocal trade due to reciprocal dumping. This problem is worsened with trade cost reductions. It can be shown that the beneficial effect due to reason 3 dominates for low transport costs, while the detrimental effect dominates for large transport costs (Brander and Krugman, 1983).
5. Intensified competition can also have a firm-selection effect, driving less-efficient firms out of the market. Recent research on firm heterogeneity and trade (Melitz, 2003) suggests this indirect effect of trade cost reduction to be always positive.
6. Trade cost reductions can trigger endogenous agglomeration according to new trade theory (Venables, 2004). Though it is often taken for granted that this causes extra gains of cost reductions, a robust proof is lacking in the literature. In fact, there are NEG models around showing that economies can tend to over- or under-agglomeration, depending on specific conditions (Pflüger and Südekum, 2008). Hence, triggering an agglomeration process may make people worse off, if it moves the economy to a state of over-agglomeration.

9.9 Conclusions

Research on infrastructure contributions to productivity are mainly carried out at the macro and meso level, making use of production functions. The common practice of using monetary valued stocks of infrastructure as an indicator of services supplied by infrastructure makes these studies less relevant for the assessment of specific projects, and also quality aspects cannot be covered in an adequate way. This calls for approaches where the services provided by infrastructure are modelled more explicitly. We review a number of recent contributions to the literature along these lines, usually implying an explicit treatment of transport network structures. A field where less progress has been made concerns the modelling of quality aspects.

Using a temporal perspective it is important to distinguish short-term effects, dominated by construction activities, and medium- to long-run effects dominated by the productivity effects. However, econometric studies cannot always distinguish between the two. Further, crowding-out effects have to be taken into account when analysing the effects on the economy. Comparing infrastructure with other types of government expenditure it appears that infrastructure has a positive impact on national production values in the large majority of cases. In this respect it ranks lower than education, but higher than other types of government expenditure.

Another direction of research where considerable progress has been made during the last decade is the modelling of spillovers and interdependencies making use of spatial econometrics. Spatial econometric approaches tend to lead to more moderate views on productivity contributions of infrastructure, but results are sometimes rather sensitive to specification.

Using welfare analysis we have shown that first-best private provision of roads is impossible, even with low exclusion costs that are technically possible in the near future. Roads tend to be increasing-returns clubs that cannot cover costs by efficient pricing. Public provision is of course not first-best either because of all forms of state failure. Hence, for choosing between private and public provision one has to weigh private versus public inefficiencies. This could not be done in this overview, but two conclusions can safely be drawn: first, the role of private provision will dramatically increase with new exclusion techniques; and second, project evaluation by public providers will still play an important role for many years to come.

Next we turned to regional evaluation and suggested the use of CGE methods. Though well founded in economic theory, the data requirements and the computational complexity prevents setting up such models in most practical cases. There is a need for short-cut methods. Partial equilibrium is such a method, leading to familiar forms of gravity models under certain convenient specifications. We have shown that measures closely related to the atheoretical classical potential measure naturally emerge as welfare measures for regional project evaluation. We believe these simplified approaches to be useful operational tools. What they do not cover is the 'wider economic effects' emerging due to inefficiencies in the economy. Principally one could cover such effects in a properly designed CGE, but our enumeration of possible effects has shown that a full account of these effects is not yet in the reach of available CGE tools. There are models with imperfect markets, but which imperfections are covered and which not is rather selective, making the net result fairly arbitrary.

Note

1. Some authors introduce a capacity measure and a capacity cost function, instead of just a capacity cost. We simplify by taking K to measure both, the capacity and its cost. This is no loss of generality: just define a unit of capacity as the amount of capacity costing one unit of money.

References

Ahmed, H. and S.M. Miller (2000), 'Crowding-out and crowding-in effects of the components of government expenditure', *Contemporary Economic Policy*, **18** (1), 124–33.

Alonso, W. (1978), 'A theory of movement', in N.M. Hansen (ed.), *Human Settlement Systems*, Cambridge: Ballinger, pp. 197–211.

Anas, A. and I. Kim (1996), 'General equilibrium models of polycentric urban land use with endogenous congestion and job agglomeration', *Journal of Urban Economics*, **40**, 217–32.

Anderson, J.E. (1979), 'A theoretical foundation for the gravity model', *American Economic Review*, **69**, 106–16.

Anderson, J.E. and E. van Wincoop (2003), 'Gravity with gravitas: a solution to the border puzzle', *American Economic Review*, **93**, 170–92.

Anderson, S.P., A. de Palma and J.-F. Thisse (1988), 'The CES and the logit: two related models of heterogeneity', *Regional Science and Urban Economics*, **18**, 155–64.

Arnott, R. and M. Kraus (2003), 'Transport economics', in R.W. Hall (ed.), *Handbook of Transportation Science*, 2nd edn, Boston, MA: Kluwer Academic Publishers, pp. 689–726.

Aschauer, D.A. (1989), 'Is public expenditure productive?', *Journal of Monetary Economics*, **23**, 177–200.

Berglas, E. (1976), 'On the theory of clubs', *American Economic Review*, **66**, 116–21.

Biehl, D. (1986), *The Contribution of Infrastructure to Regional Development*, Brussels: European Commission.

Biehl, D. (1993), *The Role of Infrastructure in Regional Development*, Mannheim: University of Mannheim.

Blum, U. (1982), 'Effects of transportation investments on regional growth', *Papers of the Regional Science Association*, **58**, 151–68.

Brander, J. and P. Krugman (1983), 'A "reciprocal dumping" model of international trade', *Journal of International Economics*, **15**, 313–21.

Briceno, C., A. Estache and N. Shafik (2004), 'Infrastructure services in developing countries: access, quality, costs and policy reform', Washington, DC: World Bank.

Bröcker, J. (2002), 'Spatial effects of European transport policy: a CGE approach', in G. Hewings, M. Sonis and D. Boyce (eds), *Trade, Networks and Hierachies: Modelling Regional and Interregional Economies*, New York: Springer, pp. 11–28.

Bröcker, J. (2004), 'Computable general equilibrium analysis in transportation economics', in D. Hensher, K.J. Button, K.E. Haynes and R. Stopher (eds), *Handbook of Transport Geography and Spatial Systems*, Amsterdam: Elsevier, pp. 269–89.

Bröcker, J., and H. Rohweder (1990), 'Barriers to international trade: methods of measurement and empirical evidence', *Annals of Regional Science*, **24**, 289–305.

Buchanan, J. (1965), 'An economic theory of clubs', *Economica*, **32**, 1–14.

Cadot, O., L.H. Röller and A. Stephan (2007), 'Contribution to productivity or pork barrel? The two faces of infrastructure investment', *Journal of Public Economics*, **90**, 1133–53.

Canning, D. (1998), 'A database of world stocks of infrastructure, 1950–95', *The World Bank Economic Review*, **12** (3), 529–47.

Costa, J.D.S., R.W. Ellson and R.C. Martin (1987), 'Public capital, regional output and development', *Journal of Regional Science*, **27**, 419–37.

de Vries, J.J., P. Nijkamp and P. Rietveld (2002), 'Estimation of Alonso's theory of movements by means of instrumental variables', *Networks and Spatial Economics*, **2**, 107–26.

Domencich, T.A. and D. McFadden (1975), *Urban Travel Demand: a Behavioral Analysis*, Amsterdam: North-Holland.

Easterly, W. and R. Levine (2001), 'What have we learned from a decade of empirical research on growth? It's not factor accumulation: stylized facts and growth models', *World Bank Economic Review*, **15**, 117–219.

Easterly, W. and S. Robelo (1993), 'Fiscal policy and economic growth: an empirical investigation', *Journal of Monetary Economics*, **32** (3), 417–58.

Fernald, J.G. (1999), 'Roads to prosperity? Assessing the link between public capital and productivity', *American Economic Review*, **89** (3), 619–638.

Flyvbjerg, B. (2003), *Mega Projects and Risk*, Cambridge: Cambridge University Press.

Flyvbjerg, B., M.K. Skamris Holm and S.L. Buhl (2005), 'How (in)accurate are demand forecasts in public works projects?', *Journal of the American Planning Association*, **71** (2), 131–46.

Forslund, U.M. and B. Johansson (1995), 'Assessing road investments: accessibility changes, cost benefit and production effects', *Annals of Regional Science*, **29**, 155–74.

Fotheringham, A.S. and M.E. O'Kelly (1989), *Spatial Interaction Models: Formulations and Applications*, Boston, MA: Kluwer Academic Publishers.
Fridstrom, L. and R. Elvik (1997), 'The barely revealed preference behind road investment priorities', *Public Choice*, **92** (1–2), 145–68.
Fukuchi, T. (1978), 'Analyse economie-politique d'un développement regional harmonise', *Collections INSEE*, **61**, 227–53.
Ginsburgh, V. and M. Keyzer (1997), *The Structure of Applied General Equilibrium*, Cambridge, MA: MIT Press.
Girard, J., H. Gruber and C. Hurst (1995), 'Increasing public investment in Europe: some practical considerations', *European Economic Review*, **39**, 731–8.
Golob, T., and M. Beckmann (1971), 'A utility model for travel forecasting', *Transportation Science*, **5**, 79–90.
Gramlich, E.M. (1994), 'Infrastructure investment: a review essay', *Journal of Economic Literature*, **32** (3), 1176–96.
Greene, W.H. (2003), *Econometric Analysis*, Washington: Pearson Education International.
Holtz-Eakin, D. and A.E. Schwartz (1995), 'Infrastructure in a structural model of economic growth', *Regional Science and Urban Economics*, **25**, 131–51.
João, M.C., J.M.C. Santos Silva and S. Tenreyro (2006), 'The log of gravity', *Review of Economics and Statistics*, **88**, 641–58.
Karlsson, C. and L. Pettersson (2005), 'Regional productivity and accessibility to knowledge and dense markets', Jonkoping: Jonkoping University.
Kelejian, H.H. and D.P. Robinson (1997), 'Infrastructure productivity estimation and its underlying econometric specifications: a sensitivity analysis', *Papers in Regional Science*, **76**, 115–31.
Kim, E. (1998), 'Economic gain and loss from public infrastructure investment', *Growth and Change*, **29**, 445–69.
Knaap, T. and J. Oosterhaven (2002), 'The welfare effects of new infrastructure: an economic geography approach to evaluating new Dutch railway links', Working Paper, Groningen University.
Knaap, T. and J. Oosterhaven (2004), 'Spatial economic impacts of transport infrastructure investments', in A. Pearman, P. Mackie and J. Nellthorp (eds), *Transport Projects, Programmes and Policies: Evaluation Needs and Capabilities*, Aldershot: Ashgate, pp. 87–105.
Kopp, A. (2005), 'Aggregate Productivity Effects of Road Investment. A Reassessment for Western Europe', 45th Congress of the European Regional Science Association, 23–27 August, Vrije Universiteit Amsterdam.
Lakshmanan, T.R., P. Nijkamp, P. Rietveld and E.T. Verhoef (2001), 'Benefits and costs of transport: classification, methodologies and policies', *Papers in Regional Science*, **80**, 139–64.
Lall, S.V. (2007), 'Infrastructure and regional growth, growth dynamics and policy relevance for India', *The Annals of Regional Science*, **41** (3), 581–601.
Lambrinidis, M., Y. Psycharis and A. Rovolis (2005), 'Regional allocation of public infrastructure investment: the case of Greece', *Regional Studies*, **39** (9), 1231–44.
Mas-Colell, A., M.D. Whinston and J.R. Green (1995), *Microeconomic Theory*, New York: Oxford University Press.
McCann, P. (2005), 'Transport costs and new economic geography', *Journal of Economic Geography*, **5**, 305–18.
Melitz, M.J. (2003), 'The impact of trade on intra-industry reallocations and aggregate industry productivity', *Econometrica*, **71**, 1695–1725.
Mera, K. (1973), 'Regional production functions and social overhead capital', *Regional and Urban Economics*, **3**, 157–86.
Mohring, H.J. (1975), 'Pricing and transportation capacity', in *Better Use of Existing Transportation Facilities*, Special Report 153, Transportation Research Board, National Research Council, Washington, DC, pp. 183–95.
Mohring, H. and M. Harwitz (1962), *Highway Benefits: An Analytic Framework*, Evanston, IL: Northwestern University Press.
Nijkamp, P., and J. Poot (2004), 'Meta-analysis of the effect of fiscal policies on long-run growth', *European Journal of Political Economy*, **20**, 91–124.
Percoco, M. (2004), 'Infrastructure and economic efficiency in Italian regions', *Networks and Spatial Economics*, **4**, 361–78.
Pereira, A.M. and J.M. Andraz (2006), 'Public investment in transportation infrastructures and regional asymmetries in Portugal', *Annals of Regional Science*, **40** (4), 803-19.
Pereira, A.M. and R. Flores de Frutos, (1999), 'Public capital accumulation and private sector performance', *Journal of Urban Economics*, **46** (2), 300–322.
Persson, T. and G. Tabellini (2002), *Political Economics. Explaining Economic Policy*, Cambridge, MA: MIT Press.
Pflüger, M. and J. Südekum (2008), 'Integration, agglomeration and welfare', *Journal of Urban Economics*, **63** (2), 544–66.
Rietveld, P. (2005), 'Transport and the environment', in T. Tietenberg and H. Folmer (eds), *The International Yearbook of Environmental and Resource Economics 2006-2007, A Survey of Current Issues*, New Horizons

in Environmental Economics Series, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, pp. 208–55.
Rietveld, P. and F. Bruinsma (1998), *Is Transport Infrastructure Effective? Transport Infrastructure and Accessibility: Impacts on the Space Economy*, Berlin: Springer-Verlag.
Romp, W. and J. De Haan (2007), 'Public capital and economic growth: a critical survey', *Perspektiven der Wirtschaftspolitik*, **8**, 6–52.
Samuelson, P.A. (1952), 'Spatial price equilibrium and linear programming', *American Economic Review*, **42**, 283–303.
Sandler, T. and J. Tschirhart (1997), 'Club theory: thirty years later', *Public Choice*, **93**, 335–55.
Seitz, H. (1995), 'The productivity and supply of urban infrastructures', *Annals of Regional Science*, **29**, 121–41.
Shoven, J.B. and J. Whalley (1984), 'Applied general equilibrium models of taxation and international trade', *Journal of Economic Literature*, **22**, 1007–51.
Small, K. (1992), 'Urban transportation economics', in J. Lesourne and H. Sonnenscheim (eds), *Fundamentals of Pure and Applied Economics*, Chur: Harwood Academic Publishers.
Stead, D. (2007), 'Transport energy efficiency in Europe: temporal and geographical trends and prospects', *Journal of Transport Geography*, **15** (5), 343–53.
Sturm, J.E., J. Jacobs and P. Groote (1999), 'Output effects of infrastructure investment in the Netherlands, 1853–1913', *Journal of Macroeconomics*, **21** (2), 355–80.
Takayama, T. and G.G. Judge (1971), *Spatial and Temporal Price and Allocation Models*, Amsterdam: North-Holland.
Tirole, J. (1988), *The Theory of Industrial Organization*, Cambridge, MA: MIT Press.
Venables, A.J. (2004), 'Evaluating urban transport improvements: cost benefit analysis in the presence of agglomeration and income taxation', CEP Discussion Paper 651, London: LSE.
Venables, A. and M. Gasiorek (1999), *The Welfare Implications of Transport Improvements in the Presence of Market Failure: The Incidence of Imperfect Competition in UK Sectors and Regions*, London: DETR.
Vickrey, W.S. (1969), 'Congestion theory and transport investment', *American Economic Review Proceedings*, **59**, 251–60.
Vijverberg, W.P.M., C. Vijverberg and J.L. Gamble (1997), 'Public capital and private productivity', *Review of Economics and Statistics*, **79**, 267–78.
Wegener, M. (2004), 'Overview of land use transport models', in D. Hensher, J. Button, K. Haynes and P. Stopher (eds), *Handbook of Transport Geography and Spatial Systems*, Amsterdam: Elsevier.
Zondag, B. (2007), 'Joint modeling of land-use transport and economy', Thesis, Technical University Delft, 24 April, TRAIL Research School, Delft.

10 Entrepreneurship and regional development

Manfred M. Fischer and Peter Nijkamp

10.1 Introduction

It is widely recognised that the region has become a fundamental basis of economic and social life. The national level of observation, though still important, is no longer the uniquely privileged point of entry to our understanding of economic development, and all the more so given the fact that the barriers between national economies are – in certain respects – breaking down, at least in Europe (Scott and Storper, 2003).

Regional economics has in the past decades made a successful attempt to uncover the complexities of the modern space-economy. It has led to important integrations of scientific perspectives, such as an integration of agglomeration theory and location theory, trade theory and welfare theory, or growth theory and entrepreneurship (including industrial organisation). The blend of rigorous economic analysis and geographical thinking has furthermore induced a bridge between two traditionally disjoint disciplines (namely, geography and economics), while this synergy has laid the foundations for innovative scientific cross-fertilisation of both a theoretical and applied nature in the important domain of regional development. The region has become a natural fruitful anchor point for an integrated perspective on the dynamics in the space-economy, such as regional development in the context of changing labour conditions, or spatial innovation in the context of metropolitan incubator conditions (see Florida, 2002).

Regions face two imperatives in a market-driven world. First, they have to be concerned with socio-economic welfare, notably employment. Job creation, an important indicator of economic growth, is central to the wealth-creating process of a regional economy. The second imperative is the ability to develop the economy. Development includes two interrelated processes: structural change and productivity improvement (Malecki, 1997a). These processes take place in a multifaceted force field.

Regional development manifests itself as a spatially uneven change in a system of regions. Regional divergence – rather than regional convergence – is a usual phenomenon that has attracted the thorough attention of both the research community and policy agencies. The standard neoclassical view of regional growth would predict that low-wage regions would acquire productive investments from high-wage regions and/or export cheap labour to these areas. The market system would then in the longer run lead to an equalisation of factor payments, so that in the final equilibrium convergence among regions would occur. In reality, this simplified model is subjected to many restrictive assumptions (full mobility, absolute cost differences, no institutional inertia, complete foresight on profitable investments, constant returns to scale), so that an equilibrium may be very hard to achieve. Regional change is ultimately the result of entrepreneurial activity in which innovations (new or improved products and processes, new management styles, locations) are key factors.

Entrepreneurship calls for risk-taking initiatives in a competitive economic environment. It encourages innovative activity and puts a region at the forefront of economic

progress. Thus, entrepreneurial culture is a prerequisite for the wealth of regions (see for example Acs, 1994; Audretsch, 2004; de Groot et al., 2004). In general, a region that hosts entrepreneurial capital and knows how to use it may be expected to be a winner in a competitive economic game. From a theoretical perspective, one might argue that regional-economic efficiency, as described by a neoclassical production function, depends critically not only on labour, capital or natural resource endowments, but also on entrepreneurial culture (including knowledge-intensive skills). The benefits of entrepreneurship for regional welfare have, in recent years, prompted much policy interest in how to favour entrepreneurship in the regional economy.

Entrepreneurship has indeed acquired central importance among the processes that affect regional economic change. Entrepreneurs are essential actors of change, and they can act to accelerate the creation, diffusion and application of new ideas. In doing so, they not only ensure the efficient use of resources but also take initiatives to exploit business opportunities. A central reason for the interest by policy-makers in entrepreneurship is its apparent capacity – based on US experience (see OECD, 1989) – to create, directly and indirectly, employment and wealth. An important indication of the significance now attached to entrepreneurship is the Organisation for Economic Co-operation and Development (OECD) study on *Fostering Entrepreneurship* to increase economic dynamism by improving the environment for entrepreneurial activity (see OECD, 1998).

This chapter makes a modest attempt to review the literature on entrepreneurship, in particular the factors that prompt entrepreneurship in the space-economy. It is not a literature on a phenomenon that has reached a mature equilibrium, but on one which is still vigorously developing. Clearly, to review such an expanding field constitutes an almost impossible task, at least as far as completeness of coverage is concerned. The chapter is organised as follows. The section that follows starts with discussing methodological and technical problems associated with research on entrepreneurship. Section 10.3 continues to derive some major factors that may explain the level of entrepreneurship, while section 10.4 provides a few observations on the spatial aspects of entrepreneurship. Section 10.5 then calls attention for entrepreneurship in a network economy, while a final section offers some retrospective and forward-looking remarks.

10.2 Definitions and measurement of entrepreneurship

Entrepreneurship is a phenomenon that takes several forms and appears in small and large firms, in new firms and established firms, in the formal and the informal economy, in legal and illegal activities, in innovative and traditional concerns, in high-risk and low-risk undertakings, and in all economic sectors (OECD, 1998). Apparently, entrepreneurship is a multifaceted phenomenon that can be viewed from different angles. Entrepreneurship has been a topic of long-standing concern in economics, but there remains little consensus on the concept of entrepreneurship (see Hébert and Link, 1989).

Different authors have stressed different facets of entrepreneurship. Schumpeter (1934), for example, emphasised the creative component. For Schumpeter the creativity of entrepreneurship lies in the ability to perceive new economic opportunities better than others do, not only in the short term as arbitrageurs, but also in the long term as fillers of innovative niches (Suarez-Villa, 1989). While in Schumpeter's concept risk-taking is not a definitional component, Knight (1921) emphasised the entrepreneur's role as dealing with risk in a context in which entrepreneurship is separable from the control of the firm. More

recently, Schultz (1980) has chosen to define entrepreneurship as the ability to deal with disequilibria rather than the ability to deal with uncertainty. Risk does not enter prominently into this concept of entrepreneurship. In his view, definitions of entrepreneurship which are uncertainty-based cannot logically relegate risk to a position of little or no importance. Finally, entrepreneurs do not work in isolation, and consequently several other economists including Piore and Sabel (1984) have stressed the network character of entrepreneurship, a new form of entrepreneurship based on innovative activities carried out in clusters of firms. A review of the conceptual and operational definitions of entrepreneurship can be found in Bögenhold (2004) (see also section 10.5).

Innovation has become a fashionable topic in modern economics, but the fundaments of this concept date back to Marshall (1890), who introduced the notion of industrial districts, in which a strong spatial concentration of (usually smaller) firms may be found and where each of these firms is specialised in one (or a few) elements of the production process of the main economic activity in the area concerned. This concentration is not only the consequence of market-driven economic and technological efficiency requirements, but is also anchored in the region's cultural, institutional and socio-economic value systems (such as trust, cooperation and social support systems). Industrial districts in general have major advantages; in particular, lower production costs, reduced transaction costs, rise in efficiency of production factors deployed and enhancement of dynamic efficiency (see Gordon and McCann, 2000; Lever, 2002; Porter, 2000). Such economic-technological clusters form the seedbed conditions for modern entrepreneurship (see Rabellotti, 1997). An extensive description and typology or regional clusters in Europe can be found in Observatory of European SMEs (2002) in which a distinction is made between regional clusters, regional innovation networks and regional innovation systems. A review of the literature on regional clusters can be found in Asheim et al. (2006).

The Schumpeterian concept of entrepreneurship remains dominant in most of the literature: the entrepreneurs as innovator and source of disequilibrium (O'Farrell, 1986; Thomas, 1987; Malecki, 1991). This corresponds to the definition of entrepreneurship proposed by the OECD (1998, p. 11):

> Entrepreneurs are essential agents of change in a market economy, and entrepreneurship fuels the drive for new economic and technological opportunities and efficient resource use . . . Growth is promoted when entrepreneurs accelerate the generation, dissemination and application of innovative ideas . . . Not only do entrepreneurs seek to exploit business opportunities by better allocating resources, they also seek entirely new possibilities for resource use.

Entrepreneurship, defined in this broad sense, is central to regional economic development. The OECD (1998, pp. 42–4) identifies three important characteristics of entrepreneurship that have emerged in the light of the above views. First, entrepreneurship involves a dynamic process in which new firms are starting up, existing firms are growing and unsuccessful ones are restructuring or closing down. This can be thought of in terms of the Schumpeterian notion of creative destruction. The dynamic structure of this process is difficult to capture empirically, but one aspect is turbulence, the rate at which businesses open and close. This notion of turbulence attempts to capture the dynamic nature of entrepreneurial activity, and has the advantage of not relying on definitions of firm size, age or growth. One widely used indicator of turbulence is firm survival rate.

Statistics of firm births may be taken from business registers. But business registers include not only data on new start-ups, but also data which do not represent births: the relocation of an existing business into another region, and the takeover of an existing business. It is difficult to identify actual start-ups, as distinct from takeovers or relocations. Firm death statistics include similar flaws: close-downs due to the sale of business and relocation. These problems make it difficult to measure survival rates accurately. Cross-regional variation in firm survival rate, moreover, could reflect differences in cyclical positions, since firm creation and destruction are sensitive to the business cycle. This complicates interregional comparisons.

A second characteristic of entrepreneurship is that – to the extent that it implies control of the process by the entrepreneur-owner – it tends to be identified with small business where the owner(s) and manager(s) are the same. One widely used measure of the extent of the combination of entrepreneurship and ownership is the self-employment or business ownership rate (Verheul et al., 2002). The term 'self-employment' refers to individuals who provide employment for themselves as business owners rather than seeking a paid job. But the entrepreneur is more than self-employed, as emphasised by Kent (1984, p. 4): 'Those who start businesses solely as an alternative to wage employment do not participate in the entrepreneurial event. Entrepreneurship requires the element of growth that leads to innovation, job-creation, and economic expansion'. Thus, not all small firms are entrepreneurships, but most entrepreneurship may be found in small firms (O'Farrell, 1986).

Finally, entrepreneurship entails innovation. This view stems from Schumpeter's (1934) suggestion that entrepreneurial innovation is the essence of capitalism and its process of creative destruction, embodied in new products, new production processes and new forms of organisation. Some technological developments, such as microelectronics and more recently biotechnology, have provided numerous opportunities for innovation and for new firms starting up and new industries to appear (Malecki, 1997a).

One measure of innovative activities is the output of the knowledge production process measured in terms of patent applications. But innovation is a phenomenon that is difficult to capture empirically. Patent-related measures have two important limitations (see Fischer et al., 2006). First, the range of patentable inventions constitutes only a subset of all research and development outcomes; and second, patenting is a strategic decision and, thus, not all patentable inventions are actually patented. As to the first limitation, purely scientific advances devoid of immediate applicability as well as incremental technological improvements – which are too trite to pass for discrete, codifiable inventions – are not patentable. The second limitation is rooted in the fact that it may be optimal for firms not to apply for patents even though their inventions would satisfy the criteria for patentability. Therefore, patentability requirements and incentives to refrain from patenting limit the measurement based on patent data. Research and development (R&D) related data, while important, relate to the input of the knowledge production process, as opposed to innovations achieved.

In general, entrepreneurs may be seen as economic change actors in an uncertain and risky business environment. Their decisions lead to spatial dynamics and are driven by dynamic efficiency objectives in which new and creative combinations of action are looked for. Under such conditions the entrepreneurial environment is excessively important: open information exchange, face-to-face interaction, presence of knowledge centres and

R&D facilities, skilled labour force, trust and solid codes of conduct, and so on (see Audretsch and Feldman, 1996; Feldman, 2000). Such factors constitute the incubation conditions for creative action in which risk-taking is an interesting option. Knowledge spillovers are then an important condition for accelerated economic development in a competitive space-economy (see for example Acs et al., 2002; de Groot et al., 2004; Nijkamp et al., 2006; Nijkamp, 2003). Especially in a major economic agglomeration with a great diversity of activities, we may observe a fluidity of information and knowledge among key actors who all benefit from the spillovers in a geographic cluster of activities. Collective learning processes and individual competitive advantages seem to reinforce each other in such a creative seedbed environment. This complex set of background conditions makes it very hard to come up with an unambiguous and conclusive (that is, measurable) definition of entrepreneurship, as the nature and creativeness of an entrepreneur is determined by the institutional context, the learning constellation of regions and Marshallian externalities.

In conclusion, there is no generally accepted definition of entrepreneurship. This reflects the fact that entrepreneurship is an elusive, multidimensional concept. It is hard to measure precisely how much entrepreneurship is taking place in a regional economy. This is difficult in part because there is no agreement on what would be appropriate and reliable indicators. Some emphasise firm start-ups and closures as an indicator of willingness to engage in risk-taking activity and capacity to innovate, and as an indicator of the ease with which resources are able to move quickly from one activity to another. Others focus on small and medium-sized enterprises where the owner(s) and manager(s) are the same. Still others associate entrepreneurship with the development of high-technology industries. None of these approaches, however, is able to provide a complete picture of the state of entrepreneurship in a regional economy. While measures of small and especially new firm development are often used as indicators of entrepreneurial activity, entrepreneurship is also critical to the maintenance of business efficiency and competitiveness in larger and longer-established businesses. The overwhelming interest nowadays in entrepreneurship is clearly induced by the competitive strategies among regions in our world, which have recognised that the presence of successful entrepreneurship and of a favourable business and innovation climate will bring high benefits to the host region.

10.3 Determinants of entrepreneurship

Regional development is a dynamic phenomenon with a permanent change in business activities. This change may be caused by innovation, by decline and by the birth and death of firms. The development of the small and medium-sized enterprise (SME) sector plays a critical role in spatial dynamics, as many forms of creative entrepreneurship are found in this sector. Clearly, the regional system (education, social support system, culture, accessibility, and so on) plays an important role in the changing conditions for entrepreneurship. Entrepreneurial adjustment patterns are thus of decisive importance for convergence or divergence patterns in regional systems. But there remains a fundamental question: which are the drivers of new business investments and new entrepreneurial modes of operation? Though in general two drivers can be distinguished, namely, new market opportunities and new consumer needs, the motivational factors of entrepreneurs call for more thorough attention.

The entrepreneurial event takes shape through the interaction of two sets of factors: personal (micro) factors and environmental (macro) factors. Much of the literature on entrepreneurship has focused on the micro factors, the characteristics of an individual to become an entrepreneur and to start a new firm. These studies focus on the role of factors such as personality, educational attainment and/or ethnic origin (Lee et al., 2004). Personality studies have found that entrepreneurship is associated with characteristics such as alertness to business opportunities, entrepreneurial vision and proactivity (see Chell et al., 1991). Research on personality, moreover, found that entrepreneurs exhibit greater individualism than non-entrepreneurs do (McGrath et al., 1992).

Monetary reward is certainly an important driver to entrepreneurship. But it is not always the prime motivation for opening up a business. Other aspects, such as the desire for independence, self-realisation, and so on often shape the entrepreneurial event. Roberts and Wainer (1971) did not find such motivational traits as these, but suggested that family background and educational attainment are most important, especially when one's father was an entrepreneur.

Studies of entrepreneurs in the United States show that the typical entrepreneur is someone in their mid-thirties to mid-forties who has worked for two or three well-established firms and decides to establish a business, often drawing directly on the skills and experience acquired in previous employment. There is a steady flow of people in the US back and forth between self-employment and salaried employment. If a business venture fails, they can reasonably easily get another job. This is much less the case in Europe because of higher unemployment, some bias against employing older workers, and the availability of early retirement.

Much of the standard research on entrepreneurship neglects the environment in which the entrepreneurial event takes place. Other more recent research, most notably Malecki (1997a), stresses the crucial role of the entrepreneurial environment for the entrepreneurial event. The meaning of the notion of environment goes here well beyond that typically used in organisation theory, but reflects the broader view including social, economic, market, political and infrastructure dimensions of the environment (Specht, 1993; Malecki, 1997a).

Roberts (1991) emphasises aspects of local culture and attributes these as critical to building a local environment that fosters entrepreneurship. Even though cultural attitudes are formed through complex processes that are not well understood, it is a generally accepted view that cultural factors affect the way in which business is done. Such factors, for example, influence the willingness to cooperate with others and may reinforce trust and personal reputation that can reduce transaction costs in doing business. Conversely, an environment characterised by mistrust may oblige entrepreneurs to spend time and money to protect against the potentially opportunistic behaviour of those with whom they work. This may deter some entrepreneurial activity. But there has been little research analysing systematically the impact of trust and mistrust on entrepreneurship.

High levels of entrepreneurial activity are often ascribed to cultural attributes. Culture, indeed, seems to play a critical role in determining the level of entrepreneurship within a region. Other things being equal, an environment in which entrepreneurship is esteemed and in which stigma does not attach to legitimate business failure will almost certainly be conducive to entrepreneurship. In the US the strong pro-entrepreneurial culture has assisted to shape institutional characteristics of the economy that facilitate

business start-ups and reward firms based on their economic efficiency. A further striking aspect of the US entrepreneurial environment is the ample availability of risk capital and generally well-functioning market mechanisms for allocating this efficiently across a wide range of size, risk and return configurations (OECD, 1998).

The key aspect of favourable entrepreneurial environments, however, is – as emphasised by Malecki (1997a) – thriving networks of entrepreneurs (see section 10.5 for further details): other firms and institutions providing capital, information and other forms of support. The theoretical notion of the milieu introduced by the GREMI group (Groupement de Recherce Européen sur les Milieux Innovateurs) epitomises these characteristics (see Maillat, 1995). Entrepreneurial development is most likely to be successful in larger urban regions, especially in metropolitan regions, where innovativeness, an entrepreneurial climate and business opportunities are relatively abundant (see Fischer and Nijkamp, 1988; Malecki, 1997a).

It should be added that knowledge-based regional innovation policies may have two constituents: (1) tailor-made support measures that enhance the micro-innovative potential of firms through the use of loans, start-up subsidies, tax credits or favourable venture capital; and (2) generic support research and R&D systems, innovation labs, university education and public–private cooperation. A further exposition on these various policies can be found in the innovation systems literature (see Lundvall, 1992; Nelson, 1993).

10.4 Spatial aspects of entrepreneurship

Entrepreneurship has in the recent past received a prominent position in economic theory, as it is increasingly recognised that entrepreneurship plays a critical role in economic growth. In contrast to traditional growth theory where technological progress and innovation was regarded as an exogenous force ('manna from heaven'), modern endogenous growth theory takes for granted that innovation and entrepreneurship are endogenous forces that are driven by various actors in the economic systems and which can be influenced by smart public policy. This new theoretical framework places much emphasis on critical success factors such as competition, vested interests, R&D, knowledge spillovers, human capital, industrial culture and entrepreneurial ability (for an overview see Capello, 2007).

In the literature on technological innovation and regional growth – following the rise of the new growth theory – three major drivers of growth were outlined: the knowledge base, innovative culture and action, and public infrastructure.

Entrepreneurship does not take place in a wonderland of no spatial dimensions, but is deeply rooted in supporting geographic locational support conditions (such as favourable urban incubation systems, venture capital support conditions, accessibility and openness of urban systems, diversity and stress conditions in the urban environment, a heterogeneous and highly skilled labour force, communication and information infrastructure, collective learning mechanisms, and so on). With the advent of the modern sophisticated communication and network structures, the action radius of entrepreneurs has significantly increased (see, for example, Reggiani and Nijkamp, 2006). Consequently, the geography of entrepreneurship and innovation has become an important field of research in modern regional economics, in which the dynamics of firms is receiving major attention.

The birth, growth, contraction and death processes of enterprises has become an important field of research in so-called firm demographics (see van Wissen, 2000). This

new field of research is concerned with the analysis of the spatio-temporal change pattern of firms from a behavioural-analytical perspective (see Nelson and Winter, 1982). Recent interesting studies in this field can be found *inter alia* in Brüderl and Schussler (1990), Siegfried and Evans (1994) and Carroll and Hannan (2000). Many studies on growth processes of firms originate from industrial economics and management disciplines (for example Stinchcombe et al., 1968; Evans, 1987; Gertler, 1988; Hayter, 1997; Caves, 1998).

The roots of this new approach can be found in the 1980s when in a period of economic recession much attention was given to the birth of new firms. From a regional economic perspective much research was undertaken on the geographical differentiation in the birth and growth process of new firms (see, for example, Keeble and Wever, 1986; Oakey, 1993; Storey, 1994; Suarez-Villa, 1996; Sutton, 1998). The predominant focus on new firm formation tended to neglect the spatio-temporal dynamics of incumbent firms, in particular the way they survive, grow or decline. From that perspective also the role of the adoption of new technology had to receive due attention (see, for example, Abernathy et al., 1983; Storper and Scott, 1989; Davelaar, 1991; Pettigrew and Whipp, 1991; Nooteboom, 1993). This has also prompted several studies on the life cycles of firms (in particular, with respect to their competitive performance, product differentiation, spatial relocation and organisational restructuring).

There are various reasons why of all types of firm dynamics, new firm formation has attracted much concern (see van Geenhuizen and Nijkamp, 1995). Perhaps most significant is the fact that new firms provide new jobs. A second reason is that new firms are often involved in the introduction of new products and processes in the market. Accordingly, they may provide a major challenge to established firms and encourage them to improve their product quality and service or to reduce prices. On the other hand, it should be recognised that newly established firms face relatively large risks, due to lack of organisational experience and cohesion. As a consequence, the death rate among start-ups is relatively high and tends to decrease over time. Many entrepreneurs appear to die at a young age. It is clear that successful new enterprises contribute significantly to the economy and employment in the region concerned. There is, however, usually a large sectoral and geographical variation among the success or survival rates of new entrepreneurs (see Acs, 1994).

Empirical research has shown that in most cases enterprises change their strategies (products, markets, and so on) in an incremental way. From historical research it appears that radical adjustments do take place, but occur rather infrequently (Mintzberg, 1978). In evolutionary economics it is emphasised that organisations develop, stabilise and follow routines. These routines may change over time, but in the short run they function as stable carriers for knowledge and experience. This causes a certain degree of 'inertia'. Related to the latter point is the core concept of search behaviour. Organisations are not invariant, but change as a result of a search for new solutions when older ones fail to work. Search behaviour follows routines, for example, based upon perceptions 'coloured' by the previous situation and biases in information processing (see also van Geenhuizen and Nijkamp, 1995).

The study of the development trajectories of individual firms from a spatio-temporal perspective is sometimes called 'company life history analysis' (see van Geenhuizen, 1993). It mainly uses a case study approach and aims to trace and explain the evolution of firms over a longer period. Particular attention is then given to entrepreneurial motives

for corporate change at the micro level. Factors to be considered are, *inter alia*, the business environment, leadership, links between strategic and operational change, human resource management and coherence in management (see also Pettigrew and Whipp, 1991). Information acquisition, for example through participation in networks of industries, is of course also an important element to be considered. In this context, the local 'milieu' may also play an important role.

It is a widely held belief that metropolitan environments offer favourable incubator conditions for creative entrepreneurship, as in this setting the conditions for proper human resource management (for example by means of specialised training and educational institutes) and labour recruitment are most favourable (see, for example, Thompson, 1968; Leone and Struyck, 1976; Pred, 1977; Davelaar, 1991; Lagendijk and Oinas, 2005). But it should be recognised that various non-metropolitan areas also offer favourable seedbed conditions for the management of corporate change. The reason is that in many non-metropolitan areas the information needs are met in localised learning mechanisms, based on a dynamic territorial interplay between actors in a coherent production system, local culture, tradition and experiences (see Camagni, 1991; Storper, 1992, 1993).

This view comes close to the one which puts a strong emphasis on the trend for localisation in less central areas where doing business is a final resort or a survival strategy. Advocates of the latter idea adhere to a vertically disintegrated and locationally fixed production, based on a shift to flexible specialisation. Some empirical evidence on non-urban seedbeds is found in high-technology regions such as Silicon Valley, Boston, the M4 Corridor, and in semi-rural areas such as the Third Italy. Although the success of economic restructuring in these regions – as a result of many high-tech start-up firms – is, without doubt, due to the pervasiveness of the trend for flexible specialisation, concomitant localisation is not sufficiently proven (see Gertler, 1988; van Geenhuizen and van der Knaap, 1994). Aside from a trend towards localisation there is a trend towards globalisation, associated with the growing influence of multinational corporations and their global networking with smaller firms (see Amin, 1993).

In the light of the previous observations it may be argued that modern entrepreneurship is based on associated skills of a varied nature. An entrepreneur is certainly an opportunity seeker, but in so doing he needs to have an open eye on a rapidly changing external environment. As a consequence, firm demography is a multidimensional field of research in which psychology, sociology, marketing, political science, economics, finance and management come together. A demographic approach to entrepreneurship may unravel various components of the spatio-temporal dynamics of both existing and new firms. In-depth case study research as advocated in company life history analysis is certainly necessary to identify motives and barriers concerning successful entrepreneurship, but there is also a clear need for more analytical comparative research leading to research synthesis and transferable lessons.

An interesting example of the latter type of research approach can be found in a recent study by Breschi (2000), who conducted a cross-sector analysis of the geography of innovative activities. Using the evolutionary concept of a technological regime he was able to identify the background factors of variations in spatial patterns of innovations, namely, knowledge base, technological opportunities, appropriability conditions and cumulativeness of technical advances. Undertaking more such studies might advance

the idea that geography counts in a modern entrepreneurial age. Cities offer important seedbed conditions for modern entrepreneurship in an open network economy, but this role is by no means exclusive. We observe at the same time local niches or shells in isolated areas which offer due protection or incubation for creative entrepreneurial abilities. Important stimulating factors may be: the presence of training and educational facilities; an open business culture; venture capital; public support; local suppliers and subcontractors; and so forth. Consequently, the geographic landscape of modern entrepreneurship is varied and calls for intensified research efforts aimed at more synthesis.

The complexity of the determinants and implications of entrepreneurial behaviour in space and time calls for sophisticated modelling efforts (see also Bertuglia et al., 1997). In the statistical analysis of entrepreneurial behaviour one may distinguish two strands of research, namely, a macro and a micro approach. In the macro approach the attention is focussed on statistical patterns and correlations between geographic location factors, entrepreneurial climate, innovative seedbed conditions, governmental support mechanisms, and so on on the one hand and entrepreneurial activity (for example investment, product choice, industrial organisation) on the other hand. Numerous studies based on aggregate figures have been performed in the past decades (see for example Audretsch et al., 2007; Santarelli, 2006; Lundstrom and Stevenson, 2005; Blanchflower et al., 2001). In the micro approach the individual motives, behavioural drivers (such as image or recognition) or cognitive determinants are analysed. This type of research is often based on survey questionnaires or interviews. There is also an abundance of literature in this field, although the spatial dimensions have been given less attention thus far (see, for example, Acs and Audretsch, 2003; Axtell et al., 2000; Baumol, 1990; Ehigie and Akpan, 2004; Campbell et al., 1996; Miron et al., 2004; Getz and Robinson, 2003).

A final remark is in order here. Entrepreneurship is not just a commercial business activity, but is often prompted by new knowledge and R&D (see also Acs, 2002; Boekema et al., 2000; Peneder, 2001). This often positions universities at the centre of creative entrepreneurship. Shane (2004) has rightly argued that academic entrepreneurship – and its related university spin-off companies – play a critical role in the commercialisation of university technology and wealth creation. The critical success factors for academic spin-offs are: the university and societal environment; the technology developed at universities; the industries favoured by these spin-offs; and the human capital involved. Thus, research and higher education are key instruments for modern entrepreneurship in knowledge-intensive regions.

10.5 Entrepreneurship and networks

A modern economy is an associative space-economy where linkages between various actors create spatial-economic externalities that are beneficial to all actors involved. Thus, modern business life is increasingly characterised by inter-actor linkage that may form complex networks. Entrepreneurship therefore also means the management of business network constellations. An interesting and rather comprehensive review of the relationship between entrepreneurship and network involvement has been given by Malecki (1997b). The local environment (including its culture, knowledge base and business attitude) appears to act as a critical success factor for new forms of entrepreneurship, a finding also obtained by Camagni (1991). Apparently, the local 'milieu' offers various types of networks which tend to encourage the 'entrepreneurial act' (see Shapero, 1984).

It should be emphasised that the chain entrepreneurship–competition–innovation–growth is not a rectilinear one. Innovation is a critical factor which functions in an open multi-actor system with concurrent phases of decisions and plan implementations, where the demand side (that is, the customer) is the driving force (see Prahalad and Ramaswamy, 2004). Innovation policy at the firm level with various risks increasingly bears a resemblance to a smart portfolio management. But in the particular case of innovation a balance has to be found between uncertain exploration and risky exploitation (March, 1991). Entrepreneurs are the foundation stones of the innovation process, as they have to create new combinations of people and products, through the creation of idea generators, of product champions, of proper support, of proper support systems and mentors, of venture mechanisms and of effective gatekeepers (see also Katz, 2003).

In the Schumpeterian view the entrepreneur is seeking for new combinations while destroying existing constellations in a creative way. This highly risk-taking behaviour, however, can be ameliorated by externalising some of the risks through participation or involvement in local or broader industrial networks. In general, the urban climate offers many possibilities for strategic network involvement, either material or virtual. In this way, the entrepreneur tends to become an organiser of change. The early urban economics literature (Hoover and Vernon, 1959) has already spelt out the great potential of urban industrial districts for creative entrepreneurship (for a review of the incubation literature, see Davelaar, 1991). Also, in the sociologically oriented writings of Jacobs (1961) we observe similar arguments. Apparently, urban modes of life create scale economies which favour the rise of new enterprises. To some extent, this idea was already propagated by Marshall (1890), who introduced the concept of industrial districts which generated an enormous economic growth potential (see also Amin and Thrift, 1992; Markusen, 1996; Paci and Usai, 2000). In general, vertical disintegration in combination with network strategies at a local level may induce a resurgence of Marshallian districts as self-contained local networks of creative economic development.

Modern information and communication technology (ICT) is a centrepiece in the rise of both local and global networks. ICT does not only induce faster and more reliable communications, but also prompts changes in firm interaction, management practice, labour acquisition and spatial structure of entrepreneurship (see Beuthe et al., 2004). In addition, ICT favours both business-to-business commerce and business-to-consumer commerce. The use of the Internet and e-commerce mean a significant and historically unprecedented rise in productivity, a phenomenon that can be ascribed to network externality theory, which explains increasing returns, first-mover advantages and coordination advantages (see for example Economides, 1996; Wigand, 1997; van Geenhuizen and Nijkamp, 2004). It is clear that creative entrepreneurship nowadays finds its roots in the modern ICT sector.

But it should be recognised that networking as a business strategy requires investments in social communication, informal bonds, training and education. To build up and to operate effectively in networks requires time and effort. Furthermore, networking may be a desirable or necessary condition, but it is by no means sufficient to ensure good entrepreneurship. And last but not least, network behaviour may also stimulate uniformity, which may contradict the entrepreneurial spirit.

Networks may, in general, relate to physical configurations (such as aviation networks, road networks, railway networks or telecommunication networks) or to virtual net-

works (such as industrial clubs, knowledge networks or information networks). Many networks may have a local character, but may also extend towards global levels. Such networks may favour industrial diversity, entrepreneurial spirit and resource mobilisation (see also Andersson, 1985; van de Ven, 1993). In general, local inter-firm networks may be seen as supporting mechanisms for new forms of creative entrepreneurship (especially among high-tech start-up firms), as such networks are a blend of openness (necessary for competition) and protection (needed for an 'infant industry'). It may be interesting to quote here the final conclusions of Malecki (1997b, p. 98): 'Thus, it is difficult for any "recipe" from one place to work when transplanted into another place, with its unique culture, traditions, capabilities, and networks'.

From the perspective of a business environment, information and knowledge is a *sine qua non* for entrepreneurial success, not only for large-scale companies but also for SMEs. Malecki and Poehling (1999), have given a very valuable review of the literature on this issue; learning-by-doing, supported by inter-firm network collaboration, enhances the competitive potential of new firm initiatives. They observe a variety of network configurations, such as suppliers or customer networks, local networks of neighbouring firms, professional networks and knowledge networks, which may all contribute to a better entrepreneurial performance. Empirical research in this area, however, is still scarce and there would be scope for more systematic comparative investigations into the knowledge drivers of modern entrepreneurship. It is certainly true that information and knowledge is an important asset in an enterprise, but the economic evaluation of such knowledge (for example as a private good or a public good with a non-rivalry character) needs to be studied more thoroughly (see Shane and Venkataraman, 2000).

An interesting illustration of the importance of local networks for new firm formation can be found in the literature on ethnic entrepreneurship (see Waldinger, 1996). Many cities in a modern industrialised world are confronted with a large influx of foreign migrants (see, for example, McManus, 1990; Borjas, 1992, 1995; Brezis and Temin, 1997; Gorter et al., 1998). The socio-economic problems involved have created an enormous tension and have prompted many policy initiatives on housing, job creation, education, and so on. But the success of such policies has not yet been impressive. The seedbed conditions for active economic participation are often weak, as a result of low levels of skill, language deficiencies, cultural gaps and stigmatisation. One of the more recent promising efforts has been to favour ethnic entrepreneurship, so that through a system of self-employment socio-cultural minorities might be able to improve their less favoured position. Ethnic entrepreneurship has different appearances, for example production for the indigenous ethnic market or low-skilled activities, but increasingly we also see an upgrading of the ethnic production sector (for example shops, software firms, consultancy).

In a recent survey study, van Delft et al. (2000) have demonstrated that the access to and use of local support networks is a critical success factor for various urban policy programmes addressing the new immigrants. Such networks may relate to socio-economic support, provision of venture capital or access to the urban community at large. The importance of social bonds and kinship relationships has also been emphasised by several other authors (for instance, Boyd, 1989; Chiswick and Miller, 1996; Borooah and Hart, 1999). In general, such networks appear to create various externalities in terms of entrepreneurial spirit, search for opportunities, self-organisation and self-education, and business information and access to local markets.

But it is noteworthy that such network connections are geared toward the geographical space in which ethnic entrepreneurs operate. It should be added that in most cities ethnic networks are not uniform, but reflect local cultures from the country of origin. Many ethnic entrepreneurs operate in volatile markets and, although network participation is needed to cope with many market uncertainties, business or social networks are usually not sufficient for an entrepreneur to survive in a competitive environment (see Barrett et al., 1996). There is a need for more thorough empirical research on the motives and performance of ethnic entrepreneurs (see also Masurel et al., 2002). The ethnic entrepreneur as a network manager is still a concept that has not become deeply rooted in the ethnic business environment.

10.6 Concluding remarks

Entrepreneurship and regional development prompt a rich variety of research questions for regional scientists. It is a domain where industrial organisation, cultural geography, location theory, business economies and technology form an intertwined nexus. From a macro or global perspective, the region is a strategic niche in a global development. But from a micro perspective, the region is shaped by innovative actions of risk-seeking entrepreneurs. Competition, trust, network organisation and public policy are ingredients for win–win situations at local level. Such elements may also offer new insights into spatial convergence debates.

There is a clear need for solid applied research on the benefits of entrepreneurship for the economic growth of regions. There is a host of anecdotal studies, but it would be a great scientific achievement to undertake a meta-analytical study on the quantitative findings in various individual studies. Such results would also offer a convincing justification of the avalanche of interest in regional entrepreneurship studies.

Our review of this complex field has not only brought to light the complex array of drivers of entrepreneurship, but has also clearly demonstrated the linkages of the theme of 'entrepreneurship and regional development' to other research domains, such as network theory, spatial externalities, cultural-behavioural theory, innovation theory and endogenous growth theory. From a dynamic entrepreneurial and regional growth theory, the interwoven connection of entrepreneurial life cycles, industrial life cycles and (multi-)regional life cycles is a fascinating research issue, not only from a theoretical viewpoint, but also from an applied modelling perspective. A particularly fascinating and policy-relevant question is then how knowledge investments and spillovers are related to dynamic spatial processes. It goes without saying that in this field a wealth of research questions are still waiting to be tackled. From this perspective, there is a great need for creative combined micro–meso–macro growth analyses at a regional level. Quantitative modelling has so far not kept pace with the research challenges in recent years.

References

Abernathy, W.J., K.B. Clark and A.M. Kantrow (1983), *Industrial Renaissance: Producing a Competitive Future for America*, New York: Basic Books.

Acs, Z. (ed.) (1994), *Regional Innovation, Knowledge and Global Change*, London: Frances Pinter.

Acs, Z. (2002), *Innovation and the Growth of Cities*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Acs, Z.J. and D.B. Audretsch (eds) (2003), *Handbook of Entrepreneurship Research*, Dordrecht: Kluwer.

Acs, Z.J., H.L.F. de Groot and P. Nijkamp (eds) (2002), *The Emergence of the Knowledge Economy*, Berlin, Heidelberg and New York: Springer.

Amin, A. (1993), 'The globalization of the economy: an erosion of regional networks?', in G. Grabher (ed.), *The Embedded Firm: On the Socio-economics of Industrial Networks*, London: Routledge, pp. 278–95.
Amin, A. and N. Thrift (1992), 'Neo-Marshallian nodes in global networks', *International Journal of Urban and Regional Research*, **16** (4), 571–87.
Andersson, A. (1985), 'Creativity on regional development', *Papers in Regional Science*, **56**, 5–20.
Asheim, B., Ph. Cooke and R. Martin (eds) (2006), *Clusters and Regional Development*, London: Routledge.
Audretsch, D.B. (2004), 'Sustaining innovation and growth: public policy support for entrepreneurship', *Industry and Innovation*, **11** (3), 167–91.
Audretsch, D.B. and M. Feldman (1996), 'Spillovers and the geography of innovation and production', *American Economic Review*, **86**, 630–40.
Audretsch, D.B., I. Grilo and A.R. Thurik (eds) (2007), *Handbook of Entrepreneurship Policy*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.
Axtell, C.M., D.J. Holman, K.L. Unsworth, T.D. Wall, P.E. Waterson and E. Harrington (2000), 'Shopfloor innovation: facilitating the suggestion and implementation of ideas', *Journal of Occupational and Organizational Psychology*, **73**, 265–85.
Barrett, G., T. Jones and D. McEvoy (1996), 'Ethnic minority business: theoretical discourse in Britain and North America', *Urban Studies*, **33** (4/5), 783–809.
Baumol, W.J. (1990), 'Entrepreneurship: productive, unproductive and destructive', *Journal of Political Economy*, **98** (5), 893–921.
Beuthe, M., V. Himanen, A. Reggiani and L. Zamparini (eds) (2004), *Transport, Development and Innovation in an Evolving World*, Berlin, Heidelberg and New York: Springer.
Blanchflower, D.G., A. Oswald and A. Stutzer (2001), 'Latent entrepreneurship across nations', *European Economic Review*, **45** (4–6), 1339–82.
Boekema, F., K. Morgan, S. Bakkers and R. Rutten (eds) (2000), *Knowledge, Innovation and Economic Growth*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.
Bögenhold, D. (2004), 'Entrepreneurship: multiple meaning and consequences', *International Journal of Entrepreneurship and Innovation Management*, **4** (1), 3–10.
Borjas, G. (1992), 'Ethnic capital and intergenerational mobility', *Quarterly Journal of Economics*, **107** (1), 123–50.
Borjas, G. (1995), 'Ethnicity, neighbourhoods and human capital externalities', *American Economic Review*, **85** (3), 365–90.
Borooah, V.K. and M. Hart (1999), 'Factors affecting self-employment among Indian and Caribbean men in Britain', *Small Business Economics*, **13**, 111–29.
Boyd, M. (1989), 'Family and personal networks in international migration: recent developments and new agendas', *International Migration Review*, **23** (3), 638–70.
Breschi, S. (2000), 'The geography of innovation: a cross-sector analysis', *Regional Studies*, **34** (3), 111–34.
Brezis, E.S. and P. Temin (eds) (1997), *Elites, Minorities and Economic Growth*, Amsterdam: North-Holland.
Brüderl, J. and R. Schussler (1990), 'Organizational mortality', *Administrative Science Quarterly*, **35**, 530–37.
Camagni, R. (1991), *Innovation Networks: Spatial Perspectives*, London: Belhaven Press.
Campbell, J.P., M.B. Gasser and F.L. Oswald (1996), 'The substantive nature of job performance variability', in K.R. Murphy (ed.), *Individual Difference and Behaviours in Organisations*, San Francisco, CA: Jossey Bass, pp. 258–99.
Capello, R. (2007), *Regional Economics*, London: Routledge.
Carroll, G.R. and M.T. Hannan (2000), *The Demography of Corporations and Industries*, Princeton, NJ: Princeton University Press.
Caves, R.E. (1998), 'Industrial organisation and new findings on the turnover and mobility of firms', *Journal of Economic Literature*, **36**, 1947–82.
Chell, E., J.M. Hawarth and S. Brearley (1991), *The Entrepreneurial Personality: Concepts, Cases and Categories*, London: Routledge.
Chiswick, B.R. and P.W. Miller (1996), 'Ethnic networks and language proficiency among immigrants', *Journal of Population Economics*, **9** (1), 19–35.
Davelaar, E.J. (1991), *Incubation and Innovation: A Spatial Perspective*, Aldershot: Ashgate.
Economides, N. (1996), 'Network externalities, complementarities and innovation', *European Journal of Political Economy*, **12**, 211–33.
Ehigie, B.O. and R.C. Akpan (2004), 'Roles of perceived leadership styles and rewards in the practice of total quality management', *Leadership and Organization Development Journal*, **25** (1), 24–40.
Evans, D. (1987), 'The relationship between firm growth, size and age', *Journal of Industrial Economics*, **35**, 567–81 .
Feldman, M.P. (2000), 'Location and innovation: the new economic geography of innovation spillovers and agglomeration', in G.R. Clark, M.P. Feldman and M.S. Gertler (eds), *The Oxford Handbook of Economic Geography*, Oxford: Oxford University Press, pp. 371–94.

Fischer, M.M. and P. Nijkamp (1988), 'The role of small firms for regional revitalization', *Annals of Regional Science*, **22** (1), 28–42.
Fischer, M.M., T. Scherngell and E. Jansenberger (2006), 'The geography of knowledge spillovers between high-technology firms in Europe: evidence from a spatial interaction modelling perspective', *Geographical Analysis*, **38** (3), 288–309.
Florida, R. (2002), *The Rise of the Creative Class*, New York: Basic Books.
Gertler, M. (1988), 'The limits of flexibility: comments on the post-Fordist vision of production and its geography', *Transactions of the Institute of British Geographers*, **17**, 410–32.
Getz, I. and A.G. Robinson (2003), 'Innovate or die: is that the fact?', *Creativity and Innovation Management*, **12** (3), 130–36.
Gordon, I.R. and P. McCann (2000), 'Industrial clusters: complexes agglomeration and/or social networks', *Urban Studies*, **37**, 513–32.
Gorter, C., P. Nijkamp and J. Poot (eds) (1998), *Crossing Borders*, Aldershot: Ashgate.
Groot, H.L.F. de, P. Nijkamp and R. Stough (eds) (2004), *Entrepreneurship and Regional Economic Development*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.
Hayter, R. (1997), *The Dynamics of Industrial Location*, Chichester: John Wiley.
Hébert, R.M. and A.N. Link (1989), 'In search of the meaning of entrepreneurship', *Small Business Economics*, **1** (1), 39–49.
Hoover, E.M. and R. Vernon (1959), *Anatomy of a Metropolis*, Cambridge, MA: Harvard University Press.
Katz, R. (2003), *The Human Side of Managing Technological Innovation*, Oxford: Oxford University Press.
Keeble, D. and E. Wever (eds) (1986), *New Firms and Regional Development in Europe*, London: Croom Helm.
Kent, C.A. (1984), 'The rediscovery of the entrepreneur', in C.A. Kent (ed.), *The Environment for Entrepreneurship*, Lexington, MA: Lexington Books, pp. 1–19.
Lagendijk, A. and P. Oinas (2005), *Proximity, Distance and Diversity*, Aldershot: Ashgate.
Lee, S.Y., R. Florida and Z.J. Acs (2004), 'Creativity and entrepreneurship: a regional analysis of new firm formation', *Regional Studies*, **38** (8), 879–89.
Leone, R.A. and R. Struyck (1976), 'The incubator hypothesis: evidence from five SMSAs', *Urban Studies*, **13**, 325–31.
Lever, W.F. (2002), 'Correlating the knowledge-base of cities with economic growth', *Urban Studies*, **39**, 859–70.
Lundstrom, A. and L. Stevenson (2005), *Entrepreneurship Policy: Theory and Practice*, Boston, MA: Kluwer.
Lundvall, B. (ed.) (1992), *National Systems of Innovation*, London: Pinter.
Maillat, D. (1995), 'Territorial dynamics, innovative milieus and regional policy', *Entrepreneurship and Regional Development*, **7**, 157–65.
Malecki, E.J. (1991), *Technology and Economic Development*, Harlow Longman Scientific & Technical.
Malecki, E.J. (1997a), *Technology and Economic Development*, Harlow: Addison Wesley Longman.
Malecki, E.J. (1997b), 'Entrepreneurs, networks, and economic development', *Advances in Entrepreneurship, Firm Emergence and Growth*, **3**, 57–118.
Malecki, E.J. and R.M. Poehling (1999), 'Extroverts and introverts: small manufacturers and their information sources', *Entrepreneurship and Regional Development*, **11**, 247–68.
March, J.G. (1991), 'Exploration and exploitation in organizational learning', *Organization Science*, **2**, 71–87.
Markusen, A. (1996), 'Sticky places in slippery space: a typology of industrial districts', *Economic Geography*, **72** (3), 293–313.
Marshall, A. (1890), *Principles of Economics*, London: Macmillan.
Masurel, E., P. Nijkamp, M. Tastan and G. Vindigni (2002), 'Motivations and performance conditions for ethnic entrepreneurship', *Growth and Change*, **33** (2), 238–60.
McGrath, R.G., I.C. Mac Millan and S. Scheinberg (1992), 'Elitists, risk-takers, and rugged individualists? An exploratory analysis of cultural differences between entrepreneurs and non-entrepreneurs', *Journal of Business Venturing*, **7**, 115–35.
McManus, W.S. (1990), 'Labour market effects of language enclaves', *Journal of Human Resources*, **25** (2), 228–52.
Mintzberg, H. (1978), 'Patterns of strategy formation', *Management Science*, **36**, 934–48.
Miron, E., M. Erez and E. Naveh (2004), 'Do personal characteristics and cultural values that promote innovation, quality, and efficiency compete complement each other?', *Journal of Organisational Behaviour*, **25** (2), 175–99.
Nelson, R. (ed.) (1993), *National Innovation Systems*, New York: Oxford University Press.
Nelson, R.R. and S.G. Winter (1982), *An Evolutionary Theory of Economic Changes*, Cambridge, MA: Harvard University Press.
Nijkamp, P. (2003), 'Entrepreneurship in a modern network economy', *Regional Studies*, **37** (4), 395–405.
Nijkamp, P., R.L. Moomaw and I. Traistaru-Siedreklag (eds) (2006), *Entrepreneurship, Investment and Spatial Dynamics*, Cheltenham, UK, and Northampton, MA, USA: Edward Elgar.

Nooteboom, B. (1993), 'Networks and transactions: do they connect?', in J. Groenewegen (ed.), *Dynamics of the Firm: Strategies of Pricing and Organisation*, Aldershot, UK and Brookfield, US: Edward Elgar, pp. 9–26.
O'Farrell, P.N. (1986), 'Entrepreneurship and regional development: some conceptual issues', *Regional Studies*, **20**, 565–74.
Oakey, R. (1993), 'High technology small firms: a more realistic evaluation of their growth potential', in C. Karlsson, B. Johannisson and D. Storey (eds), *Small Business Dynamics: International, National and Regional Perspectives*, London: Routledge, pp. 224–42.
Observatory of Regional SMEs (2002), *Regional Clusters in Europe*, European Commission, Publications Office EU, Brussels, DCJ Enterprise.
OECD (1989), *Mechanisms for Job Creation: Lessons from the United States*, Paris: Organisation for Economic Co-operation and Development.
OECD (1998), *Fostering Entrepreneurship*, Paris: Organisation for Economic Co-operation and Development.
Paci, R. and S. Usai (2000), 'Technological enclaves and industrial districts', *Regional Studies*, **34** (2), 97–114.
Peneder, M. (2001), *Entrepreneurial Competition and Industrial Location*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.
Pettigrew, A. and R. Whipp (1991), *Managing Change for Competitive Success*, Oxford: Blackwell.
Piore, M.J. and C.F. Sabel (1984), *The Second Industrial Divide: Possibilities for Prosperity*, New York: Basic Books.
Porter, M.E. (2000), 'Location, competition and economic development: local clusters in the global economy', *Economic Development Quarterly*, **14**, 15–31.
Prahalad, C.K. and V. Ramaswam (2004), *The Future of Competition*, Boston, MA: Harvard Business School Press.
Pred, A. (1977), *City-Systems in Advanced Economies*, London: Hutchinson.
Rabellotti, R. (1997), *External Economies and Cooperation in Industrial Districts*, London: Macmillan.
Reggiani, A. and P. Nijkamp (eds) (2006), *Spatial Dynamics, Networks and Modelling*, Cheltenham, UK, and Northampton, MA, USA: Edward Elgar.
Roberts, E.B. (1991), *Entrepreneurs in High Technology*, New York: Oxford University Press.
Roberts, E.B. and H.A. Wainer (1971), 'Some characteristics of technical entrepreneurs', *IEEE Transactions on Engineering Management*, **18**, 100–109.
Santarelli, E. (ed.) (2006), *Entrepreneurship, Growth, and Innovation*, Berlin, Heidelberg and New York: Springer.
Schultz, T.W. (1980), 'Investment in entrepreneurial ability', *Scandinavian Journal of Economics*, **82**, 437–48.
Schumpeter, J. (1934), *The Theory of Economic Development*, Cambridge, MA: Harvard University Press.
Scott, A.J. and M. Storper (2003), 'Regions, globalisation and development', *Regional Studies*, **37** (6/7), 579–93.
Shane, S. (2004), *Academic Entrepreneurship*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.
Shane, S. and S. Venkataraman (2000), 'The promise of entrepreneurship as a field of research', *Academic Management Review*, **96**, 98–121.
Shapero, A. (1984), 'The entrepreneurial event', in C.A. Kent (ed.), *The Environment for Entrepreneurship*, Lexington, MA: Lexington Books, pp. 21–40.
Siegfried, J.J. and L.B. Evans (1994), 'Empirical studies of entry and exit', *Review of Industrial Organisation*, **9**, 121–55.
Specht, P.H. (1993), 'Munificence and carrying capacity of the environment and organization formation', *Entrepreneurship Theory and Practice*, **17**, 77–86.
Stinchcombe, A.L., M.S. MacDill and D. Walker (1968), 'Demography of organisations', *American Journal of Sociology*, **74**, 221–9.
Storey, D.J. (1994), *Understanding the Small Business Sector*, London: Routledge.
Storper, M. (1992), 'The limits to globalization: technology district and international trade', *Economic Geography*, **68**, 60–93.
Storper, M. (1993), 'Regional "worlds" of production: learning and innovation in the technology districts of France, Italy and the USA', *Regional Studies*, **27**, 433–55.
Storper, M. and A.J. Scott (1989), 'The geographical foundations and social regulation of flexible production complexes', in J. Wolch and M. Dear (eds), *The Power of Geography: How Territory Shapes Social Life*, Boston, MA: Unwin Hyman, pp. 61–83.
Suarez-Villa, L. (1989), *The Evolution of Regional Economies: Entrepreneurship and Macroeconomic Change*, New York: Praeger.
Suarez-Villa, L. (1996), 'Innovative capacity, infrastructure and regional policy', in D.F. Batten and C. Karlsson (eds), *Infrastructure and the Complexity of Economic Development*, Berlin, Heidelberg and New York: Springer, pp. 251–70.
Sutton, J. (1998), *Market Structure and Innovation*, Cambridge, MA: MIT Press.
Thomas, M.D. (1987), 'Schumpeterian perspectives on entrepreneurship in economic development: a commentary', *Geoforum*, **18**, 173–86.

Thompson, W.R. (1968), 'Internal and external factors in the development of urban economies', in H.S. Perloff and L. Wingo (eds), *Issues in Urban Economics*, Baltimore, MD: Johns Hopkins University Press, pp. 43–62.
van Delft, H., C. Gorter and P. Nijkamp (2000), 'In search of ethnic entrepreneurship opportunities in the city: a comparative study', *Environment and Planning C*, **18** (4), 429–51.
van Geenhuizen, M. (1993), 'A longitudinal analysis of the growth of firms', PhD thesis, Erasmus University, Rotterdam.
van Geenhuizen, M. and P. Nijkamp (1995), 'A demographic approach to firm dynamics', Research Paper, Department of Economics, Free University, Amsterdam.
van Geenhuizen, M.S. and P. Nijkamp (2004), 'In search of urban futures in the e-economy', in M. Beuthe, V. Himanen, A. Reggiani and L. Zamparini (eds), *Transport, Development and Innovation in an Evolving World*, Berlin, Heidelberg and New York: Springer: pp. 64–83.
van Geenhuizen, M. and G.A. van der Knaap (1994), 'Dutch textile industry in a global economy', *Regional Studies*, **28**, 695–711.
van de Ven, A. (1993), 'The development of an infrastructure for entrepreneurship', *Journal of Business Venturing*, **8**, 211–30.
van Wissen, L.J.G. (2000), 'A micro-simulation model of firms: applications of the concept of the demography of firms', *Papers in Regional Science*, **79**, 111–34.
Verheul, I., S. Wennekers, D.B. Audretsch and R. Thurik (2002), 'An eclectic theory of entrepreneurship: policies, institutes and culture', in D.B. Audretsch, R. Thurik, I. Verheul and S. Wennekers (eds), *Entrepreneurship: Determinants and Policy in a European–US Comparison*, Berlin, Heidelberg and New York: Springer, pp. 11–81.
Waldinger, R. (1996), *Still the Promised City?*, Cambridge, MA: Harvard University Press.
Wigand, R. (1997), 'Electronic commerce', *Information Society*, **13** (1), 1–16.

PART III

DEVELOPMENT THEORIES: INNOVATION, KNOWLEDGE AND SPACE

11 Knowledge spillovers, entrepreneurship and regional development

David B. Audretsch and T. Taylor Aldridge

11.1 Introduction

The emergence of knowledge as perhaps the most decisive factor for comparative advantage also has had an impact on at least two key dimensions involving the organization of economic activity. The first involves the spatial organization of economic activity. In particular, globalization has rendered the organization of economic activity for the spatial unit of the region more important. Just as globalization has reduced the marginal cost of transmitting information and physical capital across geographic space to virtually zero, it has also shifted the comparative advantage of a high-cost *Standort*, or location, in the developed countries from being based on physical capital to being based on knowledge. This shift in the relative cost of (tacit) knowledge vis-à-vis information has been identified as increasing the value of geographic proximity. To access knowledge, locational proximity is important. Thus, a paradox of globalization is that geography has actually become more important because close spatial proximity to a knowledge can bestow competitive advantage

The second impact of globalization on the organization of economic activity involves the enterprise. While early analyses had predicted that large corporations were endowed with a competitive advantage in accessing, producing and commercializing knowledge, more recently studies have suggested that a very different organizational form – the entrepreneurial firm – has the competitive knowledge in the knowledge-based global economy.

The purpose of this chapter is to explain why the emergence of knowledge as the source of comparative advantage has rendered a shift in the organization of economic activity for both the spatial and enterprise levels. This chapter uses the lens provided by the knowledge spillover theory of entrepreneurship (Audretsch et al., 2006) to integrate both the organization of economic activity in geographic space and small firm enterprises. The knowledge spillover theory of entrepreneurship provides a focus on the generation of entrepreneurial opportunities emanating from knowledge investments by incumbent firms and public research organizations which are not fully appropriated by those incumbent enterprises. The creation of a new organization is important because it is an endogenous response to knowledge not completely and exhaustively appropriated in existing organizations. Not only does endogenous entrepreneurship serve as a conduit for knowledge spillovers, but because such knowledge spillovers tend to be spatially localized, it results in the emergence of localized entrepreneurial clusters. In the following section, the role of spatial access to access knowledge spillovers is explained. Why such knowledge spillovers do not occur automatically and may, in fact, be impeded by the knowledge filter, is explained in the third section. The role of entrepreneurship as a conduit of knowledge spillovers is explained in the fourth section. Finally, in the fifth section a summary and conclusion are provided. In particular, entrepreneurship is identified as the missing link

to regional economic growth because it provides a key mechanism facilitating the spillover and commercialization of knowledge

11.2 Knowledge spillovers and spatial proximity

An important theoretical development in the new economic geography literature is that geography may provide a relevant unit of observation within which knowledge spillovers occur. The theory of localization suggests that because geographic proximity is needed to transmit knowledge, and especially tacit knowledge, knowledge spillovers tend to be localized within a geographic region. The importance of geographic proximity for knowledge spillovers has been supported in a wave of recent empirical studies by Jaffe (1989), Jaffe et al. (1993), Acs et al. (1992, 1994), Audretsch and Feldman (1996) and Audretsch and Stephan (1996).

As it became apparent that the firm was not completely adequate as a unit of analysis for estimating the model of the knowledge production function, scholars began to look for externalities. In refocusing the model of the knowledge production to a spatial unit of observation, scholars confronted two challenges. The first one was theoretical. What was the theoretical basis for knowledge to spill over, yet at the same time be spatially within some geographic unit of observation? The second challenge involved measurement. How could knowledge spillovers be measured and identified? More than a few scholars heeded Krugman's warning (1991, p. 53) that empirical measurement of knowledge spillovers would prove to be impossible because 'knowledge flows are invisible, they leave no paper trail by which they may be measured and tracked'.

In confronting the first challenge, which involved developing a theoretical basis for geographically bounded knowledge spillovers, scholars turned to the emerging literature of the new growth theory. In explaining the increased divergence in the distribution of economic activity between countries and regions, Krugman (1991) and Romer (1986) relied on models based on increasing returns to scale in production. By increasing returns, however, Krugman and Romer did not necessarily mean at the level of observation most familiar in the industrial organization literature – the plant, or at least the firm – but rather at the level of a spatially distinguishable unit. In fact, it was assumed that the externalities across firms and even industries yield convexities in production. In particular, Krugman (1991), invoking Marshall (1920), focused on convexities arising from spillovers from: (1) a pooled labor market; (2) pecuniary externalities enabling the provision of non-traded inputs to an industry in a greater variety and at lower cost; and (3) information or technological spillovers.

Economic knowledge has long been associated with externalities. Arrow (1962) identified externalities associated with knowledge as a result of its non-exclusive and non-rival use. However, Arrow and subsequent scholars provided little insight concerning the geographic dimension of such knowledge spillovers. In particular, many authors have implicitly or explicitly assumed that knowledge externalities are so compelling that there is no reason that knowledge should stop spilling over just because of borders, such as a city limit, state line or national boundary. For example Krugman (1991), and others, did not question the existence or importance of such knowledge spillovers. In fact, they argue that such knowledge externalities are so important and forceful that there is no reason for a political boundary to limit the spatial extent of the spillover. In applying the model of the knowledge production function to spatial units of observa-

tion, theories of why knowledge externalities are spatially bounded were needed. Thus, a new theory of the development of localization was needed to explain not only that knowledge spills over but also why those spillovers decay as they move across geographic space.

Studies identifying the extent of knowledge spillovers are based on the model of the knowledge production function applied to spatial units of observation. In what is generally to be considered to be the first important study refocusing the knowledge production function, Jaffe (1989) modified the traditional approach to estimate a model specified for both spatial and product dimensions. Implicitly contained within the knowledge production function model is the assumption that innovative activity, should take place in those regions where the direct knowledge-generating inputs are the greatest, and where knowledge spillovers are the most prevalent. Jaffe (1989) dealt with the measurement problem raised by Krugman (1991) by linking the patent activity within technologies located within states to knowledge inputs located within the same spatial jurisdiction.

Empirical testing for the localization of knowledge spillovers essentially shifted the model of the knowledge production function from the unit of observation of a firm to that of a geographic unit. Jaffe (1989) found empirical evidence supporting the notion that knowledge spills over for third-party use from university research laboratories as well as from industry research and development (R&D) laboratories. Acs et al. (1992) confirmed that the knowledge production function held at a spatial unit of observation using a direct measure of innovative activity, new product introductions in the market. Feldman (1994) extended the model to consider other knowledge inputs to the commercialization of new products. The results confirmed that the knowledge production function was robust at the geographic level of analysis: the output of innovation is a function of the innovative inputs in that location.

While this literature has identified the important role that knowledge spillovers play, it provides little insight into the questions of why knowledge spills over and how it spills over. The exact links between knowledge sources and the resulting innovative output remain invisible and unknown.

One explanation was provided by the knowledge spillover theory of entrepreneurship, which suggests that the start-up of a new firm is a response to investments in knowledge and ideas by incumbent organizations that are not fully commercialized by those organizations. Thus, those contexts that are richer in knowledge will offer more entrepreneurial opportunities and therefore should also endogenously induce more entrepreneurial activity, *ceteris paribus*. By contrast, those contexts that are impoverished in knowledge will offer only meager entrepreneurial opportunities generated by knowledge spillovers, and therefore would endogenously induce less entrepreneurial activity.

Access to knowledge spillovers requires spatial proximity. While Jaffe (1989) and Audretsch and Feldman (1996) made it clear that spatial proximity is a prerequisite to accessing knowledge spillovers, they provided no insight about the actual mechanism transmitting such knowledge spillovers. As for the Romer and Lucas models (Romer 1986; Lucas, 1988, 1993), investment in new knowledge automatically generates knowledge spillovers. Their only additional insight involves the spatial dimension – knowledge spills over but the spillovers are spatially bounded.

11.3 The knowledge filter

In the Romer (1986) model of endogenous growth new technological knowledge is assumed to spill over automatically. Investment in new technological knowledge is automatically accessed by third-party firms and economic agents, resulting in the automatic spillover of knowledge. The assumption that knowledge automatically spills over is, of course, consistent with the important insight by Arrow (1962) that knowledge differs from the traditional factors of production – physical capital and (unskilled) labor – in that it is non-excludable and non-exhaustive. When the firm or economic agent uses the knowledge, it is neither exhausted nor can it be, in the absence of legal protection, precluded from use by third-party firms or other economic agents. Thus, in the spirit of the Romer model, drawing on the earlier insights about knowledge from Arrow, a large and vigorous literature has emerged obsessed with the links between intellectual property protection and the incentives for firms to invest in the creation of new knowledge through R&D and investments in human capital.

However, the preoccupation with the non-excludability and non-exhaustability of knowledge, first identified by Arrow and later carried forward and assumed in the Romer model, neglects another key insight in the original Arrow (1962) article. Arrow also identified another dimension by which knowledge differs from the traditional factors of production. This other dimension involves the greater degree of uncertainty, higher extent of asymmetries, and greater cost of transacting new ideas. The expected value of any new idea is highly uncertain, and as Arrow pointed out, has a much greater variance than would be associated with the deployment of traditional factors of production. After all, there is relative certainty about what a standard piece of capital equipment can do, or what an (unskilled) worker can contribute to a mass-production assembly line. By contrast, Arrow emphasized that when it comes to innovation, there is uncertainty about whether the new product can be produced, how it can be produced, and whether sufficient demand for that visualized new product might actually materialize.

In addition, new ideas are typically associated with considerable asymmetries. In order to evaluate a proposal concerning a new idea in, say, nanotechnology, the decision-maker might need to have not only a PhD in nanotechnology, but also a specialization in the exact scientific area. Such divergences in education, background and experience can result in a divergence in the expected value of a new project or the variance in outcomes anticipated from pursuing that new idea, both of which can lead to divergences in the recognition and evaluation of opportunities across economic agents and decision-making hierarchies. Such divergences in the valuation of new ideas will become greater if the new idea is not consistent with the core competence and technological trajectory of the incumbent firm.

Thus, because of the conditions inherent in knowledge – high uncertainty, asymmetries and transaction costs – decision-making hierarchies can reach the decision not to pursue and try to commercialize new ideas that individual economic agents, or groups or teams of economic agents think are potentially valuable and should be pursued. The basic conditions characterizing new knowledge, combined with a broad spectrum of institutions, rules and regulations, impose what Audretsch et al. (2006) term the 'knowledge filter'. The knowledge filter is the gap between new knowledge and what Arrow (1962) referred to as economic knowledge or commercialized knowledge. The greater is the knowledge filter, the more pronounced is this gap between new knowledge and new economic, or

commercialized, knowledge. The knowledge filter is a consequence of the basic conditions inherent in new knowledge.

11.4 Entrepreneurship as a conduit of knowledge spillovers

The knowledge filter is a consequence of the basic conditions inherent in new knowledge. Similarly, it is the knowledge filter that creates the opportunity for entrepreneurship in the knowledge spillover theory of entrepreneurship. According to this theory, opportunities for entrepreneurship are the duality of the knowledge filter. The higher is the knowledge filter, the greater are the divergences in the valuation of new ideas across economic agents and the decision-making hierarchies of incumbent firms. Entrepreneurial opportunities are generated not just by investments in new knowledge and ideas, but in the propensity for only a distinct subset of those opportunities to be fully pursued by incumbent firms.

The discrepancy in organizational context between the organizations creating opportunities and those exploiting the opportunities that seemingly contradicted Griliches' model (1979) of the firm knowledge production function was resolved by Audretsch (1995), who introduced the 'knowledge spillover theory of entrepreneurship':

> the findings challenge an assumption implicit to the knowledge production function: that firms exist exogenously and then endogenously seek out and apply knowledge inputs to generate innovative output. It is the knowledge in the possession of economic agents that is exogenous, and in an effort to appropriate the returns from that knowledge, the spillover of knowledge from its producing entity involves endogenously creating a new firm. (pp. 179–80)

What is the source of this entrepreneurial opportunity that endogenously generates the start-up of new firms? The answer seemed to be through the spillover of knowledge that creates the opportunities for the start-up of a new firm:

> How are these small and frequently new firms able to generate innovative output when undertaken a generally negligible amount of investment into knowledge-generating inputs, such as R&D? One answer is apparently through exploiting knowledge created by expenditures on research in universities and on R&D in large corporations. (p. 179)

The empirical evidence supporting the knowledge spillover theory of entrepreneurship was provided from analyzing variations in start-up rates across different industries reflecting different underlying knowledge contexts. In particular, those industries with a greater investment in new knowledge also exhibited higher start-up rates, while those industries with less investment in new knowledge exhibited lower start-up rates, which was interpreted as a conduit transmitting knowledge spillovers.

Thus, compelling evidence was provided suggesting that entrepreneurship is an endogenous response to opportunities created but not exploited by the incumbent firms. This involved an organizational dimension involving the mechanism transmitting knowledge spillovers – the start-up of new firms. In addition, Jaffe (1989), Audretsch and Feldman (1996) and Audretsch and Stephan (1996) provided evidence concerning the spatial dimension of knowledge spillovers. In particular their findings suggested that knowledge spillovers are geographically bounded and localized within spatial proximity to the knowledge source. None of these studies, however, identified the actual mechanisms which actually transmit the knowledge spillover; rather, the spillovers were implicitly

assumed to exist automatically (or fall like manna from heaven), but only within a geographically bounded spatial area.

While much has been made of the key role played by the recognition of opportunities in the cognitive process underlying the decision to become an entrepreneur, relatively little has been written about the actual source of such entrepreneurial opportunities. The knowledge spillover theory of entrepreneurship identifies one source of entrepreneurial opportunities – new knowledge and ideas. In particular, the knowledge spillover theory of entrepreneurship posits that it is new knowledge and ideas created in one context, but left uncommercialized or not vigorously pursued by the source actually creating those ideas, such as a research laboratory in a large corporation or research undertaken by a university, that serves as the source of knowledge generating entrepreneurial opportunities. Thus, in this view, one mechanism for recognizing new opportunities and actually implementing them by starting a new firm involves the spillover of knowledge. The organization creating the opportunities is not the same organization that exploits the opportunities. If the exploitation of those opportunities by the entrepreneur does not involve full payment to the firm for producing those opportunities, such as a license or royalty, then the entrepreneurial act of starting a new firm serves as a mechanism for knowledge spillovers.

Thus, the knowledge spillover theory of entrepreneurship shifts the fundamental decision-making unit of observation in the model of the knowledge production function away from exogenously assumed firms to individuals, such as scientists, engineers or other knowledge workers – agents with endowments of new economic knowledge. As Audretsch (1995) pointed out, when the lens is shifted away from the firm to the individual as the relevant unit of observation, the appropriability issue remains, but the question becomes: 'How can economic agents with a given endowment of new knowledge best appropriate the returns from that knowledge?' If the scientist or engineer can pursue the new idea within the organizational structure of the firm developing the knowledge, and appropriate roughly the expected value of that knowledge, the worker has no reason to leave the firm. On the other hand, if the scientist places a greater value on his ideas than do the decision-making bureaucracy of the incumbent firm, they may choose to start a new firm to appropriate the value of his knowledge.

In the knowledge spillover theory of entrepreneurship the knowledge production function is actually reversed. The knowledge is exogenous and embodied in a worker. The firm is created endogenously in the worker's effort to appropriate the value of their knowledge through innovative activity. Typically an employee from an established large corporation, often a scientist or engineer working in a research laboratory, will have an idea for an invention and ultimately for an innovation. Accompanying this potential innovation is an expected net return from the new product. The inventor would expect to be compensated for their potential innovation accordingly. If the company has a different, presumably lower, valuation of the potential innovation, it may decide either not to pursue its development, or that it merits a lower level of compensation than that expected by the employee.

In either case, the employee will weigh the alternative of starting their own firm. If the gap in the expected return accruing from the potential innovation between the inventor and the corporate decision-maker is sufficiently large, and if the cost of starting a new firm is sufficiently low, the employee may decide to leave the large corporation and

establish a new enterprise. Since the knowledge was generated in the established corporation, the new start-up is considered to be a spin-off from the existing firm. Such start-ups typically do not have direct access to a large R&D laboratory. Rather, the entrepreneurial opportunity emanates from the knowledge and experience accrued from the R&D laboratories with the entrepreneurs' previous employers. Thus the knowledge spillover view of entrepreneurship is actually a theory of endogenous entrepreneurship, where entrepreneurship is an endogenous response to opportunities created by investments in new knowledge that are not commercialized because of the knowledge filter.

As already discussed, a vigorous literature has identified that knowledge spillovers are greater in the presence of knowledge investments. Just as Jaffe (1989) and Audretsch and Feldman (1996) show, those regions with high knowledge investments experience a high level of knowledge spillovers, and those regions with a low amount of knowledge investments experience a low level of knowledge spillovers, since there is less knowledge to be spilled over.

Thus, as a result of the knowledge filter, entrepreneurship becomes central to generating economic growth by serving as a conduit for knowledge spillovers. The process involved in recognizing new opportunities emanating from investments in knowledge and new ideas, and attempting to commercialize those new ideas through the process of starting a new firm, is the mechanism by which at least some knowledge spillovers occur. In the counterfactual situation, that is, in the absence of such entrepreneurship, the new ideas would not be pursued, and the knowledge would not be commercialized. Thus, entrepreneurs serve as an important mechanism in the process of economic growth. An entrepreneur is an agent of change, who recognizes an opportunity, in this case generated by the creation of knowledge not adequately pursued (in the view of the entrepreneur) by incumbent organizations, and ultimately chooses to act on that opportunity by starting a new firm.

As investments in new knowledge increase, entrepreneurial opportunities will also increase. Contexts where new knowledge plays an important role are associated with a greater degree of uncertainty and asymmetries across economic agents evaluating the potential value of new ideas. Thus, a context involving more new knowledge will also impose a greater divergence in the evaluation of that knowledge across economic agents, resulting in a greater variance in the outcome expected from commercializing those ideas. It is this gap in the valuation of new ideas across economic agents, or between economic agents and decision-making hierarchies of incumbent enterprises, that creates the entrepreneurial opportunity.

By serving as a conduit for the spillover of knowledge that otherwise might not have been commercialized, entrepreneurship provides a missing link to economic growth. Because the spillover of knowledge tends to be localized within the spatial context of geographically bounded regions, entrepreneurship becomes an important vehicle in regional clusters by which (regional) knowledge spills over and becomes transmitted into (regional) growth.

The knowledge spillover theory of entrepreneurship analogously suggests that, *ceteris paribus*, entrepreneurial activity will tend to be greater in contexts where investments in new knowledge are relatively high, since the new firm will be started from knowledge that has spilled over from the source actually producing that new knowledge. A paucity of new ideas in an impoverished knowledge context will generate only limited entrepreneurial

opportunities. By contrast, in a high knowledge context, new ideas will generate entrepreneurial opportunities by exploiting (potential) spillovers of that knowledge. Thus, the knowledge spillover view of entrepreneurship provides a clear link that entrepreneurial activity will result from investments in new knowledge and that entrepreneurial activity will be spatially localized within close geographic proximity to the knowledge source.

The 'endogeneous entrepreneurship hypothesis' involves the organizational interdependency between entrepreneurial start-ups and incumbent organizations investing in the creation of new knowledge (Audretsch et al., 2006; Audretsch, 2007). A second hypothesis emerging from the knowledge spillover theory of entrepreneurship, the 'localizational hypothesis', has to do with the location of the entrepreneurial activity and the key role that regional clusters play. Since we have just identified one such mechanism by which knowledge spillovers are transmitted – the start-up of a new firm – it follows that knowledge spillover entrepreneurship is also spatially bounded in that local access is required to access the knowledge facilitating the entrepreneurial start-up. According to the localization hypothesis, knowledge spillover entrepreneurship will tend to be spatially located within close geographic proximity to the source of knowledge actually producing that knowledge. Thus, in order to access spillovers, new firm start-ups will tend to locate close to knowledge sources, such as universities.

Systematic empirical evidence consistent with the knowledge spillover theory of entrepreneurship has been provided by Audretsch et al. (2006) and Acs et al. (2004). Both studies find that entrepreneurship rates tend to be greater in the context of greater investments in new knowledge.

11.5 Conclusions

Along with globalization has come a shift in the comparative advantage of developed countries towards knowledge-based economic activity. This shift towards a knowledge-based economy has not left the organization of economic activity unchanged. Rather, this chapter has identified two key dimensions involving the organization of economic activity that have changed in virtually every developed country. The first involves the spatial dimension of economic activity. As knowledge becomes more important, so too has the spatial concentration of knowledge activities, which facilitates the spillover of knowledge.

However, this chapter has explained why the spillover of investments in new knowledge is by no means automatic. Rather, the knowledge filter can impede the spillover and commercialization of knowledge. By serving as a conduit for knowledge spillovers, entrepreneurship is the missing link between investments in new knowledge and economic growth. Thus, the spillover theory of entrepreneurship provides not just an explanation of why entrepreneurship has become more prevalent as the factor of knowledge has emerged as a crucial source for comparative advantage, but also why entrepreneurship plays a vital role in generating economic growth. Entrepreneurship is an important mechanism permeating the knowledge filter to facilitate the spillover of knowledge and ultimately to generate economic growth.

Using the lens provided by the knowledge spillover theory of entrepreneurship, this chapter has explained why location is the underlying organizational context for entrepreneurship. Just as knowledge spillovers have been found to be spatially bounded, entrepreneurship has been shown to be an important conduit by which that knowledge spills over. Taken together, these two organizational units form the basis for entrepreneurial clusters.

A generation ago, the entrepreneurial firm within the context of regional clusters did not seem to be prominent in the public policy approach to enhancing growth and creating employment. For example, in advocating a new public policy approach to promote growth and international competitiveness at the European level, Servan-Schreiber warned of the 'American Challenge' in the form of the 'dynamism, organization, innovation, and boldness that characterize the giant American corporations' (1968, p. 153). Because giant corporations were considered to be the engine of growth and innovation, Servan-Schreiber advocated the 'creation of large industrial units which are able both in size and management to compete with the American giants' (1968, p. 159). According to Servan-Schreiber (1968, p. 159):

> The first problem of an industrial policy for Europe consists in choosing 50 to 100 firms which, once they are large enough, would be the most likely to become world leaders of modern technology in their fields. At the moment we are simply letting industry be gradually destroyed by the superior power of American corporations.

Ironically, the 1988 Cecchini Report identified the gains from European integration as largely accruing from increases in scale economies. However, the more recent insights concerning the role of entrepreneurship and regional clusters have become a focal point in the debate to foster growth and employment. For example, in the Lisbon Accord of 2000, the European Commission made a formal commitment to becoming the entrepreneurship and knowledge leader in the world in order to foster economic growth and prosperity throughout the European Union. Similarly, as Bresnahan and Gambardella (2004, p. 1) point out:

> Clusters of high-tech industry, such as Silicon Valley, have received a great deal of attention from scholars and in the public policy arena. National economic growth can be fueled by development of such clusters. In the United States the long boom of the 1980s and 1990s was largely driven by growth in the information technology industries in a few regional clusters. Innovation and entrepreneurship can be supported by a number of mechanisms operating within a cluster, such as easy access to capital, knowledge about technology and markets, and collaborators.

Similarly, Wallsten (2004, p. 229) suggests that: 'Policy makers around the world are anxious to find tools that will help their regions emulate the success of Silicon Valley and create new centers of innovation and high technology.' Little is actually known about which specific instruments will best serve public policy in creating knowledge-based entrepreneurial clusters. What has become clearer is that these two fundamental changes in the organization of economic activity, one at the spatial level and the other at the enterprise level, hold the key to generating economic growth, jobs and competitiveness in a globalized economy.

References

Acs, Z.J., D.B. Audretsch and M.P. Feldman (1992), 'Real effects of academic research: comment', *American Economic Review*, **82**, 363–7.

Acs, Z.J., D.B. Audretsch and M.P. Feldman (1994), 'R&D spillovers and innovative activity', *Managerial and Decision Economics*, **15**, 131–8.

Acs, Zoltan J., David B. Audretsch, Pontus Braunerhjelm and Bo Carlsson (2004), 'The missing link: the knowledge filter and entrepreneurship in endogenous growth', Centre for Economic Policy Research (CEPR) Discussion Paper.

Arrow, K. (1962), 'Economic welfare and the allocation of resources for invention', in *The Rate and Direction of Inventive Activity*, Princeton, NJ: Princeton University Press, pp. 609–26.
Audretsch, D.B. (1995), *Innovation and Industry Evolution*, Cambridge, MA: MIT Press.
Audretsch, D. (2007), *The Entrepreneurial Society*, New York: Oxford University Press.
Audretsch, D. and M. Feldman (1996), 'R&D spillovers and the geography of innovation and production', *American Economic Review*, **86**, 630–40.
Audretsch, David B. and Paula E. Stephan (1996), 'Company–scientist locational links: the case of biotechnology', *American Economic Review*, **86** (3), 641–652.
Audretsch, David B., Max Keilbach and Erik Lehmann (2006), *Entrepreneurship and Economic Growth*, New York: Oxford University Press.
Bresnahan, T. and A. Gambardella (2004), *Building High-Tech Clusters: Silicon Valley and Beyond*, Cambridge: Cambridge University Press.
Feldman, M.P. (1994), 'Knowledge complementarity and innovation', *Small Business Economics*, **6**, 363–72.
Glaeser, E., H. Kallal, J. Scheinkman and A. Shleifer (1992), 'Growth in cities', *Journal of Political Economy*, **100**, 1126–52.
Griliches, Zvi (1979), 'Issues in assessing the contribution of R&D to productivity growth', *Bell Journal of Economics*, **10** (Spring), 92–116.
Jaffe, Adam B. (1989), 'Real effects of academic research', *American Economic Review*, **79** (5), 957–70.
Jaffe, A., M. Trajtenberg and R. Henderson (1993), 'Geographic localization of knowledge spillovers as evidenced by patent citations', *Quarterly Journal of Economics*, **63**, 577–98.
Jovanovic, Boyan (2001), 'New technology and the small firm', *Small Business Economics*, **16** (1), 53–5.
Krugman, P. (1991), *Geography and Trade*, Cambridge, MA: MIT Press.
Lucas, R. (1988), 'On the mechanics of economic development', *Journal of Monetary Economics*, **22**, 3–42.
Lucas, R. (1993), 'Making a miracle', *Econometrica*, **61**, 251–72.
Marshall, A. (1920), *Principles of Economics*, 8th edn, London: Macmillan.
Romer, P. (1986), 'Increasing returns and long-run growth', *Journal of Political Economy*, **94**, 1002–37.
Servan-Schreiber, J. (1968), *The American Challenge*, London: Hamish Hamilton.
Solow, R. (1956), 'A contribution to theory of economic growth', *Quarterly Journal of Economics*, **70**, 65–94.
Wallsten, S. (2004), 'The role of government in regional technology development: the effects of public venture capital and science parks', in T. Bresnahan and A. Gambardella (eds), *Building High-Tech Clusters: Silicon Valley and Beyond*, Cambridge: Cambridge University Press, pp. 229–79.

12 R&D spillovers and regional growth

Daria Denti

12.1 Introduction

In many, if not all, developed countries there is now a lively debate about research and development (R&D), policy and its influences on country economic performance. The source for this debate goes back to the end of the Cold War, when the national-defence rationale for a variety of well-established governmental practices to support R&D broke down. Its place has promptly been taken by economic arguments, such as international competitiveness and gross domestic product (GDP) growth (Stokes, 1997). The shift in the *raison d'être* of public support towards innovation has granted continuity in subsidization but it has also disclosed new questions about the proper way to design R&D policy. Sceptical authors argue that R&D policy based on economic arguments misses real background as it hinges on assumptions which are not supported by the data and that, moreover, it has a negative effect on the economy as public support towards R&D displaces private resources (Kealy and Rudenski, 1998). Others argue that the public-good features of R&D, among which are R&D spillovers, are such that private agents fail to capture all the economic benefits from R&D investment and therefore they invest at a level which is lower than socially desirable. As a consequence, the government must intervene either through subsidization to increase the private willingness to invest in R&D or by directly investing in some amount of R&D (David, 1997).

Many contributions within economic literature have investigated the linkages between R&D and economic performance of countries: in these works R&D is pursued by private agents which are driven by economic incentives. Technological change is endogenous as it depends on the allocation of the economy's resources and it affects the growth rate. Moreover, technological progress is the proposed outcome of an economic activity – investment in R&D – and not an unintended by-product of other activities, such as learning-by-doing.

The analysis of an economy with endogenous technological change has been carried out both at a theoretical and at an applied level, by constructing endogenous growth frameworks where R&D activity is intentionally pursued by firms and by testing both their assumptions and predictions through econometric models. The main results of this bulk of literature support the key role of R&D investments in promoting economic growth and, at the same time, these works offer a large variety of questions looking for an answer. Some of the most important issues concern: the effects of publicly funded R&D on industrial productivity growth; the linkage between a country's R&D investment level and international competition; the best mix of basic and applied R&D expenditure for the economy; the optimal allocation of private and public resources to R&D; and the effectiveness of different sets of fiscal tools adopted to promote growth through technology creation. This chapter will provide a review of many influential and path-breaking works tackling the aforementioned questions.

As both R&D and technology are endowed with some peculiarities leading to some kind of market failure, often it turns out to be optimal to support them in order to overcome the low incentive that private agents would otherwise have to invest in R&D. Anyhow, as R&D is a manifold activity and its output has powerful and wide effects, the literature offers a rich menu of policy advice on how to design the optimal supporting scheme: depending on the features it is actually clothed with, it affects growth and welfare in a particular way that can be captured by the proper model of growth.

The main conclusion is that there is no general policy advice on how to deal optimally with R&D. The variety of proper fiscal tools depends heavily on the menu of R&D spillovers that are influencing the economy. Positive spillovers call for public support, but it may also be the case that R&D exerts negative externality effects.

From an applied point of view, there is wide support towards the positive effects of R&D spillovers on economic growth. Moreover, there are many interesting works exploring in more detail the structure of the linkage between the two variables. The focus is mainly on regional aspects (geographical effects, social effects, and so on), convergence and decomposition of the R&D activity. Results from these empirical analyses show that opening up the R&D black box and the trajectories of its influence provide important insights.

We will also briefly introduce the up-to-date issue relating R&D to environmental preservation. Embedding environmental issues in R&D-based growth models or, more generally, dealing with pollution along with innovation, generally means focusing on R&D explicitly pursued to abate pollution. This vision, although widely popular, need not encompass all the pathways through which R&D, production and pollution are interrelated. In fact, even if it is undoubtedly true that firms do R&D to increase their profits, it may be the case that some of their R&D exerts some positive effects on pollution abatement as an unintended by-product. These features are channelled by spillovers, initiating new and unexpected patterns and processes with positive effects on environmental care.

The remainder of the chapter is organized as follows. First, we describe R&D as an economic activity and technology as a commodity, and we outline their main features and the possible effects that they exert on the economy. The focus will be strongly on the effects of the wide menu of possible spillovers that R&D is acknowledged to affect the economy with. We will also discuss some developments towards the disentanglement of R&D activity and to the analysis of the peculiar features of each component. Then, as R&D spillovers are among the key elements driving endogenous technological change in developed countries and have thus become a keystone upon which many policy agendas are designed, we review some of the most relevant recent contributions to endogenous growth theory dealing with inserting some aspects of R&D and technology inside a growth model. To this respect, we will consider some of the most influential and recent approaches to the embedment of R&D spillovers in a growing economy and the way each approach shapes growth. Finally, we survey empirical works testing the significance of the theoretical findings of the models.

12.2 R&D spills over

R&D and technology

R&D is an economic activity and technology is the output of this activity. Both R&D and technology are endowed with some interesting characteristics that have been deeply analysed by economists as they exert substantial influences on the economic system.

Technology as a commodity

The economic effects of technology are due to its quasi public good nature (for example Arrow, 1962; Grossman and Helpman, 1991). Technology is non-rival and partially non-excludable. Moreover, some other features must be added: the incremental cost of an additional user is zero; the extensive use of knowledge does not deplete the commodity, as it may also increase it; and accumulation of knowledge does not have physical bounds.

Non-rivalry and partial excludability of technology imply that private economic agents cannot find the proper incentive to invest in R&D as they will fail to appropriate the returns. An intellectual property rights system helps to foster the private incentive to invest in R&D, by increasing the level of private returns from an R&D investment.

Besides creating market failure, non-excludability of technology determines another key feature the consequences of which on the economic performance cannot be neglected: as innovators cannot fully prevent other agents from using the knowledge that they have created, then existing knowledge becomes a free input in the production of new knowledge. Conversely, it may also be the case that as knowledge increases, it becomes more difficult to find a new idea; if this is the case, then existing knowledge still flows throughout the economy, but it has a negative effect on new idea creation.Technological spillovers from R&D activities are a key element in explaining growth patterns: they imply that innovation can be a self-fulfilling process, or that at least its creation benefits from a free input given by existing knowledge spilling over.

The structure of R&D processes

R&D is an economic activity made of different steps. Simply looking at the acronym, it is clear to see that it contains two abbreviations: research and development. R&D literature identifies three main steps characterizing the process of new design creation: basic (or fundamental or blue-sky) research; applied research and development; and it provides a definition to identify them within R&D. Obviously, each stage deals with knowledge creation, and differences are mainly due to the aim the efforts are devoted to. Basic research is defined as: 'systematic study directed towards greater knowledge or understanding of the fundamental aspects of phenomena and of observable facts without specific application towards procedures and products in mind' (Eisenman et al., 2002; NSF,[1] 2004). Applied research is defined as 'knowledge necessary for determining the means by which a recognized and specific need may be met and research projects which represent investigations directed to discovery of new scientific knowledge and which have specific commercial objectives with respect to either products or processes' (Eisenman et al., 2002; NSF, 2004). Development is defined as: 'a systematic application of knowledge towards the production of useful materials, devices, systems and methods, to meet specific requirements' (Eisenman et al., 2002; NSF, 2004).A look at US data about R&D shows that private agents are indeed performing the three components, although it is commonly thought that basic research is carried on exclusively at the public level (NSF, 2004).

Starting in the 1960s and 1970s, industrial innovation in the US was identified with some corporate entities, which were pursuing far-sighted research. Recruitment of researchers was aimed at attracting the most able people, who were provided with a great deal of latitude in performing R&D. The most famous companies adopting this strategy were General Electric, IBM, Bell and Xerox, and their scientific achievements have been recognized by several Nobel prizes (Auerswald et al., 2005). Nowadays, companies' support for

'blue-sky' research has changed, as the majority of firms tends to prefer investing in short-term R&D. However, it is important to notice that there are important exceptions: in some industries, such as electronics and chemistry, firms invest significant amounts of money in fundamental research as they reckon this to be the most suitable strategy for long-term survival (Auerswald et al., 2005); a key example is given by the Google R&D strategy encouraging engineers and scientists working in the company to spend a fraction of their working schedule on whatever research project strikes their fancy.[2]

Having identified R&D components and assessed their weight, it is now time to discuss both their economic characteristics and how these activities are organized inside the R&D black box. With respect to the economic features of R&D components, both economic theory and empirical works have highlighted meaningful distinctions between basic research and development along a variety of dimensions (for example Audretsch et al., 2002; Nelson, 1959; Pavitt, 2001). An increasing amount of literature on innovation, technology and R&D management acknowledges the huge impact of basic research – also privately performed – in shaping the patterns of innovation and its returns (Branscomb and Auerswald, 2001). These contributions have identified several distinguishing features which we briefly present.

Basic research is the R&D activity the output of which is the most likely to fail to be directly economically exploitable in order to produce new intermediate goods.[3] The same elements implying that not all the designs developed by basic research efforts are economically exploitable, while the designs resulting from development efforts are more likely to be so, also entail that basic research is more likely to generate breakthrough innovations than development activity (Nelson, 1959; Theis and Horn, 2003). Basic research does not have any precise goods, process or prototype to work on. Its aim is mainly the exploration of the unknown. Obviously, new, breakthrough innovations are more likely to come out from new understanding of something that was previously unknown than from improvement and enhancement of what is already known. If we look at the literature on R&D, we see much emphasis devoted to the role of basic research in generating breakthrough innovations[4] (Audretsch et al., 2002; Theis and Horn, 2003).

Even when a basic research design is economically exploitable, it usually needs further efforts to become suitable for the production of an intermediate good (Nelson, 1959; Auerswald et. al., 2005). Basic research generates positive and significant spillovers affecting the economy across sectors, whereas spillovers associated to development activity are generally weak and do not propagate across different sectors (for example Lichtemberg and Siegel, 1991; Kesteloot and Veugelers, 1995; Funk, 2002). Both literature and evidence suggest that we have to distinguish basic research-intensive activities from development-intensive ones along the following lines: (1) development-intensive activities, when compared to basic research-intensive activities, give innovators a higher probability of getting positive pay-offs, since they generate mainly designs which are useful in economic terms; (2) pay-offs generated by economic exploitation of development-intensive designs have, in turn, a higher probability of lasting for a shorter time horizon due to close substitution effects; (3) basic research-intensive designs, when economically exploitable, usually need further research efforts to be suitable for the production of an intermediate good; (4) basic research generates the strongest and most pervasive positive spillovers. Points (1) to (3) affect the structure of pay-offs from R&D efforts. Point (3) affects the structure of the research process.

It is widely documented that basic research has been playing a key role in the US political agenda since the 1940s: basic research is considered fundamental to gain major achievements in many different fields, therefore keeping the US's leading position as an exporter of goods and services (US Office for Management Budget, 2004), and it is acknowledged to be both necessary and sufficient for technical progress (Stokes, 1997; Pavitt, 2001). This political vision has actually determined the continuous flow of public funding for basic research at both academic and firm level that we see in US data and that is now under debate in many countries.

Reaching the organizational issue, in the literature there are essentially two benchmark models (Audretsch et al., 2002; Stokes, 1997). The 'linear paradigm' states that basic research is the first stage of R&D, its aim being to broaden the frontier of knowledge without any specific practical end, whereas the aim of both applied research and development is the transformation of the output of the previous steps into designs to be used in the production of goods and services. The process is linear as there is no feedback from subsequent steps to the previous one. Conversely, the 'dynamic paradigm' takes into account the fact that certain basic research processes are motivated by technological problems highlighted by applied research and development activities. Thus, even if basic research is still the preliminary step in the research process, it is inspired by specific problems in goods and services production or improvement, and the feedback linkage among the different R&D activities compose a web of spillovers. Case studies, historical and anecdotal evidence show that both ways operate (Stokes, 1997).

Externalities and the components of an R&D process

The different degrees of generality between basic research and development determine differences in the pervasiveness of spillovers associated to each R&D component. The relationship between degree of generality and pervasiveness and strength of spillovers across the economy is a positive one.

Another interesting point refers to the spillovers occurring within R&D, that is, feedback (from development to basic research) and the positive externality effect from knowledge (from basic research to development). This issue must be analysed considering the benchmark structures for an R&D process: the linear paradigm and the dynamic paradigm.

If we add spillover trajectories to these models, then the analysis becomes more complicated. There are two ways through which spillovers from R&D components may influence the multi-stage R&D process. The linear paradigm entails a unique direction for the flow of knowledge, from basic research to development with no feedback. So, only research externalities may affect the process by influencing development. The dynamic paradigm broadens the possible trajectories for knowledge, so that feedbacks from development to basic research are also allowed (Stokes, 1997). Figure 12.1 summarizes the structure of an R&D process and the direction of spillovers.

12.3 R&D, technological externality and growth

Traditional growth models where technological change is assumed to be exogenous suffer from some shortcomings, due to their inadequacy to explain the determinants of long-run economic growth and to their failure to account for technical progress deliberately pursued by economic agents as acknowledged by data on industrial organization. Therefore, increasing efforts have been devoted to endogenize technology and innovation processes.

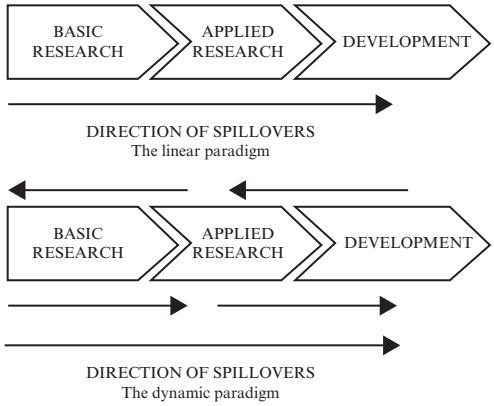


Figure 12.1 R&D structures models

Starting from the 1990s, many contributions have provided growth models where innovation is an endogenous activity intentionally carried on by firms. The development of these frameworks was made possible thanks to important developments in industrial organization theory with respect to the treatment of investments in innovation (Gancia and Zilibotti, 2005). Empirical evidence shows that firms are actually investing their money in R&D, thus it is important to understand which are the driving forces behind this behaviour and which are the suitable ways to formalize them. Industrial organization provides the answer: private incentives to invest in R&D are due to the market power granted to innovators by the introduction of a new or improved good or process. Obviously, market power itself is preserved as long as technology is at least partially excludable; excludability with respect to knowledge is accomplished through copyrighting. Therefore, as long as an intellectual property right system exists and is enforced, there is a positive incentive to invest in R&D, as the inventor will enjoy a sufficient degree of protection with respect to the commercial exploitation of his/her design to produce commodities. Protection with respect to commercial exploitation is formalized through either royalty payments by licensees or monopolistic pay-offs from the direct in-house production of either the commodity or the process produced out of the design. Protection can be modelled in many ways: complete and never ending, stochastic, partial, and so on.

Having defined the way investments in R&D take place, it is still to be determined how this activity is able to generate growth endogenously. In this respect, there are mainly three approaches, and all deal with R&D spillovers.

Some authors argue that R&D efforts by firms generate new varieties of good and processes. Therefore innovation implies an expansion in the set of available commodities. This expansion offsets the tendencies for diminishing returns typical of both labour and capital in the production of output. In this way there is room for endogenous growth. Assuming that innovation benefits the economy by increasing the set of variety of goods hinges on the idea that the availability of more goods, either for final consumption or as intermediate inputs, raises the material well-being of people. This can happen through two channels: consumers' taste for variety and increase in productivity due to a larger set of available tools.

Other authors support another vision: R&D is pursued by firms to improve the quality of existing commodities. Under this scenario, it is the increasing quality of goods that offsets decreasing returns to traditional inputs, thus creating endogenous growth.

Within growth literature, the former viewpoint is called 'horizontal innovation literature' and the latter 'Schumpeterian growth literature'. The two perspectives depict two complementary attributes of R&D: innovation takes place both through the creation of new varieties of goods and processes and through the improvement of existing varieties. To account for complementarity, some recent attempts tackle endogenous growth models where R&D accomplishes both tasks at the same time.

Technological spillovers exert a key function in each case, although the channels through which they operate and also the dimension of the set of spillovers taken into account change considerably.

Horizontal innovation and R&D spillovers

Seminal contributions to the theory of horizontal innovation date back to the late 1980s and early 1990s. In those years some authors developed dynamic models of economic growth with monopolistic competition and innovation motivated by profits (for example Judd, 1985; Grossman and Helpman, 1989; Romer, 1990). Among them, Romer was the first to design fully a set-up where private R&D efforts driven by monopolistic pay-offs from innovation generate growth endogenously. In this set-up technology spillovers are accounted for and they exert a key role in shaping equilibrium allocation and growth. A simplified version of his model is described below.[5]

The baseline model To capture the main insights of the model we can abstract from investments in physical capital and consider a simplified environment. There are three types of agents in the model. Households maximize utility subject to their budget constraint. They hold shares of intermediate sector firms, supply labour and invest in new ideas. Final good producers hire labour and intermediate goods and combine them to produce a final good, which is sold at unit price. This final good serves different purposes: consumption, and input for intermediate good production. R&D activity is privately performed and relies on labour as costly input.

An important assumption is that innovation generates an intertemporal technological externality: the productivity of scientists increases with the stock of knowledge. This assumption can be interpreted according to Newton's statement: 'If I have seen further, it is by standing on the shoulders of giants': researchers benefit from accessing the stock of existing knowledge, thereby obtaining inspiration for new designs.

Producers of final goods have access to a production technology combining a number of intermediate inputs and labour to produce final output, which is then sold in the market at unit price. Formally, $Y = (\int_0^N x_j^{\alpha} dj) L_Y^{1-\alpha}$, where $0 < \alpha < 1$. Final good sector aggregates in a Cobb–Douglas fashion two costly inputs: intermediate goods, x_j, and labour, L_Y. x_j is the employment of the j-th type of intermediate good and N is the total number of varieties of intermediates in the economy. N corresponds to the total stock of designs available at time t. The price of Y is taken as the numeraire. Each intermediate good producer holds a patent which grants the exclusive right to produce a specific variety of intermediate good. Every patent allowing for a new variety grants perpetual monopolistic profits to the producer. We assume that an intermediate good, once invented, costs one unit of Y to

produce and it is used in the production of the final good forever. Therefore, obsolescence is ruled out. This assumption is indeed quite strong and it should refer only to breakthrough innovations, although in this family of models it is applied to all kind of designs.

Technological advancements take place through an expansion in the variety of intermediate goods, formally through an increase in N. The same increase in N drives endogenous growth, by offsetting decreasing returns to labour and intermediate goods in final good production.

Expansion in the size of varieties is driven by new firms wishing to enter intermediate good production. In fact, these firms must invest in research first. Firms face a two-stage decision process. First, they decide whether to enter or not. Entrants will invest in R&D if the market value of the firm producing the new variety of intermediate good is at least as large as the R&D expenditure they have to bear to start the firm. Then, they decide the optimal price at which to sell their new intermediate goods to final good sector firms. The two-stage problem is solved backward. The market value of a new intermediate good firm is given by $V = \int_t^{\infty} e^{-\int_t^s r(\tau)d\tau}\pi(s)ds$, where π is the instantaneous profits from intermediate good production. The market values for a new intermediate good firm is given by $V=(\pi/r)$. Free entry implies that V cannot exceed the entry cost. R&D cost is determined by R&D firms' profit-maximization problem and it corresponds to the price of a new patent. This price is determined by R&D firms according to their technology. We determine the price considering the maximization problem of the i-th firm in the R&D sector.

To generate a final blueprint that can be sold in the market to entrant firms, the R&D firm i must undertake a costly activity described by the following technology $b_i = (1/\eta)NL_{Ni}$, where b_i corresponds to the single new blueprint produced by R&D firm i. L_{Ni} are scientists employed by R&D firm i-th, η is an exogenous productivity parameter and N is the existing stock of knowledge exerting a positive spillover effect on new design creation.

Having analysed the problem of firm i, we need to reach the aggregate level of the R&D sector, to determine how knowledge evolves in the economy. At the aggregate level there are N identical innovators, therefore new designs evolve according to $N = (1/\eta)NL_N$.

Finally, population is constant, and equal to L. Households maximize utility over an infinite horizon. They supply labour inelastically to both final good production and knowledge creation. Their objective function is CES in consumption. Households receive a wage rate on labour and returns from assets. They discount the future at rate ρ. The consumption plan they set when maximizing utility subject to the constraints satisfies the standard Euler equation.

The final good sector is perfectly competitive. Firms are price-takers and they maximize profits using intermediate goods and workers as costly inputs. The first-order conditions of this maximization problem give the optimal level for wage and intermediate good price. The latter becomes the indirect demand for type j intermediate goods and it is used by monopolistic producers to solve their maximization problem. Using these results, we can find the present discounted value for starting a new firm V. The discounted flow of pay-offs from entry must be equal to the entry cost to have positive entry in the economy. The entry cost is given by the price of a patent that comes from R&D firm profit maximization. Solution of the profit maximization yields

$\hat{p} = \eta(1-\alpha)\alpha^{(2\alpha/1-\alpha)}N/N$. This is the expression for the marginal cost of innovation. It is straightforward that it does not depend on the knowledge stock as N disappears. In fact, labour productivity and, hence, the equilibrium wage grow linearly with N. But also, productivity of scientists increases with N, due to the intertemporal technological spillover. The two effects cancel out. Note that, without the technological externality, the entry cost would be increasing in the stock of knowledge and this would imply that growth would eventually cease.

Free entry implies the equilibrium expression for the rate of return, which must be equal to another equilibrium expression for the rate of return, that has been determined exploiting the assumption that all variables in the economy grow at the same, constant rate (balanced growth path). Thus, we reach the equilibrium expression for the growth rate $\gamma = (1/\eta)(\alpha\overline{L} - \eta\rho/\alpha + \sigma)$.

The equilibrium allocations and the resulting growth rate determined in the decentralized economy are not Pareto optimal. Market failures are due to monopolistic competition and to technological spillovers. With respect to the latter, as private agents fail to internalize the social effects of R&D investments because they are not captured in the payoffs from innovation, the decentralized setting experiences a level of investments in R&D that is lower than socially optimal. To solve for this distortion, it is optimal to set a subsidy to R&D spending. This policy advice has provided the theoretical support to R&D-promoting policy, such as the US tax credit.

Schumpeterian growth with spillovers

Technological progress has been modelled also as enhancements in the quality level of existing types of good. This approach is the complement of the horizontal innovation concept, as technological progress happens through both an improvement in the quality of existing goods and the introduction of new goods. The complementary aspect of the two approaches has been recognized and recent contributions have, in fact, developed models where R&D efforts are directed towards both directions. These works will be discussed in the next section, whereas here we focus on the baseline model of quality improvement and the role exerted by technological spillovers. The seminal contributions to this strand of the literature on endogenous growth are due to Aghion and Howitt (1999) and Grossman and Helpman (1991).

The baseline model[6] The economy is made of three sectors: final good producers, consumers and R&D performers. As before, the final good is produced using a continuum of intermediate goods and labour,

$$Y = \left(\sum_{j=1}^{N} (q^{k_j} x_j)^{\alpha} dj \right) L^{1-\alpha}.$$

But now the number of varieties is fixed, N, and each intermediate good j is associated with its quality level, q^{k_j}: $q^{k_j}x_j$ indicates the quality-adjusted amount of good j used in final good production. Each intermediate good in the continuum has a quality level and these levels are ranked along a quality ladder whose rungs are spaced proportionately at interval $q > 1$. The starting quality level for each good is normalized to unity, then it jumps to q, then to q^2 and so on and so forth.

Increasing the quality level of an intermediate good is a costly R&D activity. Therefore, technological progress takes place through increasing k_j. Note that improvements are sequential and it is not possible to skip any rung.

As before, private agents need the proper incentive to invest in R&D and start production. In this respect, the assumption about complete and never-ending protection of the production of the intermediate good granted by the design still holds. However, in this set-up there is obsolescence and, as a consequence, replacement occurs. Therefore the monopolistic profits generated by intermediate good production are granted as long as the intermediate good is not replaced by one of the same type, but is endowed with superior quality. Then, even if the patent law still grants exclusivity of commercial exploitation of the good, the good itself is not required in production any more. An important assumption is that the final good is produced using exclusively intermediate goods of the highest quality available. This obviously implies that only intermediate goods of the highest possible quality are produced. An important assumption states that incumbents – that are intermediate good producers – cannot discover the next rung of their intermediate good. In other words, improved-quality of existing varieties are generated only by entrants.

Each intermediate good, no matter its quality level, costs one unit of final good to be produced. R&D firms face the same two-stage decision process described in the baseline model of horizontal innovation. Entrants will invest in R&D if the market value of the firm producing the new variety of intermediate good is at least as large as the R&D expenditure they have to bear to start the firm. Demand for x_j is used by the intermediate good producer to solve its profit-maximization problem, which yields the optimal price and the quality-adjusted optimal size for the good. Using these intermediate results, we determine profits from intermediate good production, which are an increasing function of quality.

As there are N industries and vertical innovation takes place in each of them, substituting the equilibrium value for x_j inside final output technology, it is possible to identify the aggregate quality index,

$$Q = \sum_{j=1}^{N} q^{k_j} \frac{\alpha}{1-\alpha}.$$

Evaluation of the present discounted value of starting a new firm considers the time span through which the owner of the design for good j of quality k_j retains the highest-quality design for x_j, enjoying the consequent monopolistic pay-offs. As soon as a competitor develops x_j of quality k_{j+1}, profits drop to zero, as the good is replaced. The time span is not exogenously given, as it is determined by the endogenous R&D efforts borne by competitors. In fact, R&D efforts determine the probability that a superior quality is created. Calling $t_{k_{j+1}}$ the moment in which the improvement appears and t_{k_j} the moment in which x_j of quality k_j appeared, we can determine the present discounted value of the firm for the standpoint of the inventor of quality level k_j. This is a random variable whose expected value $E(V)$ depends on a Poisson process governing the probability of erosion of monopolistic profits, which is summarized by s_{k_j}. The next step is determining the relationship between the probability governing the Poisson process and the R&D effort. The assumption done here states that s_{k_j} depends linearly on the aggregate R&D effort in industry j, proxied by total R&D investments, and on the current level of technology,

k_j: $s_{k_j} = Z_{k_j}\phi(k_j)$ where Z_{k_j} stands for aggregate R&D investments in industry j at current technological level k_j and $\phi(k_j)$ is the dependence on the current level of technology.

Again, there is free entry in the business of being an inventor. Moreover, we assume that innovators care about $E(V)$, taking the rate of return as deterministic. Free entry implies $s_{k_j}E(V_{k_{j+1}}) = Z_{k_j}$, which can be simplified using intermediate results and imposing that R&D efforts must be positive in equilibrium $r + s_{k_{j+1}} = \phi(k_j)\tilde{\pi}_{k_{j+1}}$, where is intermediate good profits.

By assuming that $\phi(k_j)$ depends negatively on the current level of knowledge, it is possible to simplify the analytical treatment of the model. It can be easily shown that in this economy all variables grow at the same constant rate, which is driven by the growth rate of Q. To close the model it is necessary to determine the equilibrium expression for this growth rate. The quality index changes whenever new quality improvements arise, an event that happens after positive R&D effort, with probability s per period of time. We assume that the probability of success is the same throughout the whole economy and, if the economy is large enough, we can assume that Q is differentiable and its growth rate is non-stochastic. Finally, exploiting uniqueness of growth rate and the Euler equation it is possible to identify the decentralized outcome:

$$\gamma = \left(q^{\frac{\alpha}{1-\alpha}} - 1\right)\left[\frac{(1-\alpha)}{\alpha}\alpha^{\frac{2}{1-\alpha}}L}{\zeta} - \rho\right] \Big/ \left[1 + \sigma\left(q^{\frac{\alpha}{1-\alpha}} - 1\right)\right].$$

This value for the growth rate and the implicit rate of return contains two distortions due to R&D. There are the positive technological spillovers exerted by the current technological level on the subsequent R&D efforts. This effect is at work also in the baseline model of horizontal innovation. Private agents cannot capture it, as they see technology benefits ending as the next innovation replaces the existing one. So, R&D is underperformed.

Then, there is another externality, driven by the fact that private agents fail to internalize the loss caused by an innovation to the previous incumbent. This negative externality is called 'business stealing', and it implies that there is too much R&D performed in the economy.

Therefore, Schumpeterian models of endogenous technological change introduce a new source of externality coming from R&D, while keeping the positive intertemporal R&D spillovers. Overall it may be that either the positive or the negative prevails, making the decentralized outcome respectively lower or higher than socially optimal.

Towards a broader menu of technology spillovers

The Jones critique The strongest critique of the baseline models concerns the displayed scale effect played by population size on the growth rate, as it is not supported by data (Jones, 1995). As the critique has a strong empirical support, it is important to check what happens to equilibrium allocations, growth and policy advice when the set-up is modified in a way that removes the scale effect. The role of technological spillovers is crucial in discarding the scale effect, although there might be important consequences on growth.

Consider an economy which is summarized by the same equations of the baseline of the horizontal innovation model described before, but the R&D technology. Hence, the

law for R&D accumulation changes and there will be important consequences for growth prediction.

With respect to R&D, there is a different way of thinking about technology spillovers. It is acknowledged that some major innovations benefit productivity of scientists afterwards, but at the same time it is assumed that the more knowledge accumulates, the more difficult it becomes to discover a new variety. Formally, the positive spillovers from technology exhibit diminishing returns (Jones, 1995).

Moreover, R&D generates other side-effects: duplication of ideas and overlapping. These are negative externalities from technology; therefore this set-up boosts the trajectories through which technology affects the economic system. Duplication and overlapping reduce the productivity of scientists (Jones, 1995).

Overall, R&D technology and the law of motion for knowledge change as follows: $b_i = (1/\eta)N^{\phi}L_{Ni}l_N^{\lambda-1}$ and $\dot{N} = (1/\eta)N^{\phi}L_N l_N^{\lambda-1}$. N^{ϕ}, with $0 < \phi < 1$, accounts for diminishing returns from previous knowledge. $l_N^{\lambda-1}$stands for negative externality from duplication and $0 < \lambda \leqslant 1$.[7] In equilibrium, $l_N = L_{N.}$

In this set up, if all variables grow at the same rate in equilibrium (Balanced Growth Path), than the steady state growth rate is determined by the law of motion for knowledge, $\gamma = (1/\eta)N^{\phi-1}L_N^{\lambda}$.

This expression can be differentiated, recalling that, along the balanced growth path, the growth rate of knowledge is constant by definition: $\gamma = (\lambda\gamma_L)/(1-\phi)$, where γ_L stands for the growth rate of population. If we assume that population grows at a positive rate which is exogenously determined and equal to n, then the growth rate is independent from the size of the population.

These results are to be ascribed to the menu of externality effects played by R&D on its own creation process.

However, growth is now semi-endogenous: the growth rate of endogenous variables depends on the growth rate of population, which is exogenous. If population is not growing, then the economy is not growing either. Dependence of growth on an exogenous variable takes place notwithstanding endogenous technological change.

Another key finding refers to government tax policy. The decentralized solution coincides with the social optimum, therefore subsidizing R&D is neutral with respect to growth.

The expansion in the externality effects of technology has determined the neutrality of R&D policy with respect to growth, but at the same time it also changes static allocation of resources in equilibrium. Differently from the baseline model, technological spillovers are both positive and negative. The main consequences are that the level of scientists' employment determined in the decentralized setting fails to account for technology externalities, and that due to the presence of the negative duplication effect and of diminishing returns from past knowledge externality, it is quite likely that private agents are overinvesting in R&D, and if they are not, it is a consequence of monopolistic competition. This finding contrasts heavily with public support to private R&D, as it calls for policy to compensate for monopolistic competition only.

Negative spillovers from R&D To some extent it is plausible to consider that R&D becomes progressively more difficult over time. This accounts for assuming that the most obvious ideas are discovered first, making it harder to make new discoveries afterwards. This idea has been explored by some authors (for example Barro and Sala-i-Martin,

2005; Segestrom, 1998) and it finds empirical support with respect to some countries and industries (Segestrom, 1998).

The easiest set-up is analysed by Barro and Sala-i-Martin (2005): the model coincides with the baseline model of horizontal innovation with the sole exception of R&D technology and the consequent law of motion for R&D: $b_i = (1/\phi N^{\sigma})L_{Ni}$ and $\dot{N} = (1/\eta N^{\sigma})\, L_N$, where both σ and ϕ are positive parameters. It is clear that now the stock of existing knowledge exerts a negative effect on new design creation. This change affects the patent price accordingly. Accounting for this in the free entry condition implies that the interest rate is no longer constant. An important consequence of negative technological spillovers causes the rate of return to be influenced both negatively by the negative externality because it reduces the productivity of scientists, but also positively by knowledge through the increase in the value of the firm, a value that increases because subsequent innovations are more difficult to be created and there is no obsolescence. In this economy the variables grow at different rates and, in the long run, there is semi-endogenous growth.

A more complicated model accounting for the same negative externality effect from R&D has been developed by Segestrom (2000). R&D determines quality improvements of existing varieties, but discovering new improvements becomes more difficult as knowledge accumulates. Each inventor displaces a previous incumbent and starts earning monopolistic profits. There are several spillovers from R&D: negative intertemporal technological spillovers; a negative business-stealing effect typical of Schumpeterian models of growth; and positive spillovers accruing to consumers whose utility depends positively on quality. To account for the negative intertemporal spillover effect from R&D, an R&D difficulty index is introduced and it is assumed that this index increases as the R&D efforts, $\dot{X}(i,t) = \eta I(i,t)X(i,t)$, where η, $\eta>0$, is an exogenous productivity parameter, $I(i,t)$ is the probability of success of an R&D investment in industry i at time t and $X(.)$ is the R&D difficulty index. Then, differently from the baseline model of Schumpeterian growth, the probability of a successful R&D effort depends linearly on R&D effort, but negatively on the R&D difficulty index, $I(i,t) = AL_i(i,t)/X(i,t)$, where $L_i(i,t)$ stands for the aggregate R&D effort in industry i, measured by total scientists employment in the industry. A is a given technology parameter.

Given these assumptions, the economy converges towards an equilibrium of semi-endogenous growth where the resources allocated to R&D do not match with the first-best due to the different spillovers associated to R&D: consumers benefit (positive), business stealing (negative) and intertemporal technological effect (negative). Subsidizing R&D becomes Pareto improving only under some specific circumstances. Moreover, it may indeed be the case that R&D should be optimally taxed. The general finding states that the higher the quality of the improvement, the lower the subsidy, as the negative externality effects prevail over the positive one.

Multiple R&D-sector economy and technology trajectories

Besides criticisms of some of the results of the benchmark models, it is acknowledged that they make a pivotal contribution to the understanding of endogenous technological change. Thus, many works have embedded further features to account for observed facts about private R&D investments and management.

Multiple R&D sectors and market structures A focus that appears to be particularly relevant refers to the role of market structure in shaping firms' decisions about R&D investments.

High-tech industries are characterized by several features deserving attention: there are some firms – the technology champions – that are pushing out the technological frontier and there is enough evidence showing that the relationship between market structure and R&D investments is important. A formal treatment of these facts introduces some important analytical features; one refers to the process of R&D creation and it considers increasing returns to scale related to knowledge accumulation that are internal to the firm; the other relates the R&D strategy chosen by a firm to the level of competitiveness of the industry. The strand of growth literature that investigates the linkages between market concentration, firms, size and endogenous technological change is quite recent, but it has provided important insights towards a deeper understanding of the possible trajectories through which R&D and technology affect the economy.

The seminal paper in this field deals with a two-R&D-sector economy where oligopolistic producers of intermediate goods accumulate labour-augmenting knowledge in order to reduce production costs, offer lower prices and thereby expand sales (Peretto, 2003, 1999a). In this set-up, different R&D activities are introduced: there is 'in-house' R&D carried on by firms to produce cost-reducing innovations with respect to their own production, and there is also entrepreneurial R&D which is performed to develop a new product and enter production. This way of modelling R&D recognizes local incremental technological progress pursued by each firm through internal costly activities. Knowledge produced through in-house R&D is firm-specific, whereas entrepreneurial R&D has broader scope, as it aims at new variety creation.

Technological spillovers are also at work in this framework. They increase the productivity of scientists developing in-house R&D. Although this activity is local in its target, it benefits from a positive externality effect coming from the stock of public knowledge. Moreover, there is an implicit technology spillover from incumbent firms to new entrants as the latter have to hire incumbent workers to set up production and these workers are endowed with positive externality effects coming from the stock of public knowledge from their previous jobs. The main properties of the decentralized equilibrium link entry, market concentration and growth. Although these seminal contributions deliver meaningful insights to the relationship between market structure, R&D strategies and growth, they do not give a central role to technology spillovers, as the uniquely meaningful spillovers are those between incumbents and entrants, but they cease in equilibrium. It is then interesting to try to embed more trajectories for R&D spillovers in the set-up.

To overcome this shortcoming, it seems quite reasonable to start by considering technological spillovers between R&D activities, that is, trajectories between in-house R&D and entrepreneurial R&D, to account for feedback that improved management of production of existing varieties may exert on new breakthrough innovations. One of the main consequences of the introduction of inter- and intra-sector spillovers between two R&D activities is semi-endogenous growth. In particular, endogenous growth happens under very specific and stringent conditions, whereas semi-endogenous growth is a more general finding (Li, 2000).

We consider an economy with two types of R&D activity: in-house R&D performed to improve the quality of an existing variety and entrepreneurial R&D to set up a new

variety of intermediate good. Intermediate goods are then used to produce final goods according to $Y = [\int_0^N (q_j x_j)^{\alpha} dj]^{1/\alpha}$, where q_j stands for the quality of intermediate good x_j. Intermediate goods are produced using labour through a one-to-one technology. If symmetry applies, we have that the aggregate pool of intermediate good is produced using the aggregate fraction of total labour devoted to production, $(1-b)L$. So, $(1-b)L = Nx$. Using this result inside the final output technology and using symmetry also for quality, it is straightforward to see that output growth is driven by technology advancements in both direction: expansion of varieties, N, and improvement of qualities, q:

$$\gamma_Y = \left(\frac{1-\alpha}{\alpha}\right)\gamma_N + \gamma_q$$

Then, new varieties are created using labour and benefiting both from existing knowledge about varieties and knowledge about quality improvements $\dot{N} = mbLK_N/q$, where mbL is the fraction of labour devoted to R&D reserved for new variety creation. R&D externalities on new variety creation go through K_N and q. The former depends both on the existing pool of ideas on varieties, N, and on the existing pool of knowledge about quality, q.[8] The latter is due to the assumption that the more quality evolves, the more difficult it is to create a new variety of the same quality. Formally, $K_N = N^{\phi_N} q^{\delta_N}, \phi, \delta > 0$.

Analogously, quality improvements accumulate according to $\dot{q} = (1-m)bLK_q/N$, where R&D externalities propagate only through K_q. As before, K_q depends both on N and q: $K_q = N^{\phi_q} q^{\delta_q}, \phi, \delta > 0$. Note that each R&D activity benefits from both intrasector and inter-sector spillovers, a feature that was neglected in the previous models. Substituting for the expressions for $K_h, h = N, q$, and considering the growth rate, we find: $\gamma_N = mbL/(q^{1-\delta_N} N^{1-\phi_N})$ and $\gamma_q = (1-m)bL/(q^{1-\delta_q} N^{1-\phi_q})$.

Along the balanced growth path, both growth rates must be constant. This property, together with the fact that population is assumed to grow at a constant positive rate exogenously given by λ, determine a system in λ, γ_N and γ_q, whose solution implies that the growth rate of final output is driven by population growth and parameters governing the extent of technological spillovers. Both inter- and intra-sector spillovers are meaningful in shaping growth, as both δ_h and ϕ_h, $h = N, q$, influence growth. Dependence of R&D growth and final output growth on population exogenous growth implies that growth is semi-endogenous.

The result can be generalized in a setting where the number of dimensions of technological progress is increased. Moreover, increasing the menu of R&D activities makes the requirements for endogenous growth even more stringent.

Growth becomes endogenous only when the technological spillovers exerted by each R&D activity are the same, both inter- and intra-sector. However, assuming that knowledge externalities are the same, both inter- and intra-sector, does not meet empirical evidence, where it is clearly demonstrated that spillovers are generally stronger intra-sector.

In a recent contribution, Strulik (2005) introduces human capital into the framework developed by Li. Human capital is a factor that can be accumulated, and in equilibrium it drives growth. Therefore, the addition of human capital makes growth fully endogenous.

R&D in a multi-industry economy Another interesting approach to allow for the rich set of technological spillover trajectories recognized by empirical evidence hinges on an economy characterized by many industries which differ in the type of output they

generate and where economic incentives determine in which industry to invest. This approach analyses the forces driving the direction of technical change, hinging on data supporting the claim that technological change is not neutral: technological progress tends to be more labour-augmenting than capital-augmenting; however technical change typically favours unskilled labour over skilled labour (Gancia and Zilibotti, 2005). Tackling these topics offers another opportunity to consider inter-sector technological spillovers. In particular, it is possible to consider inter- and intra-industry technological spillovers.

The seminal contribution in the field is due to Acemoglu and Zilibotti (2001), who have developed a set-up able to investigate the influence of technology spillovers in determining the direction of technical change.

The set-up mainly extends the baseline model of horizontal innovation in a two-industry dimension. The two industries are labelled L and Z and they produce different final goods, Y_L and Y_Z, using industry-specific intermediate goods and industry-specific fixed factors (skilled labour for industry Z and unskilled labour for industry L). Each industry has local monopolists producing industry-specific intermediate goods after having either purchased or produced a specific blueprint. The two final goods are aggregated and the resulting commodity is used for consumption and as an input to produce intermediate goods in both industries.

Differently from the two approaches reviewed above, in this set-up each industry has a unique R&D activity and technological advance in each industry takes place exclusively along an expansion in the number of varieties. Here multiplicity of R&D happens through multiplicity of industries.

The two final goods are produced according to the following technologies $Y_L = (1/\alpha)(\int_0^{N_L} x_{Lj}^{\alpha} dj) L^{1-\alpha}$ and $Y_Z = (1/\alpha)(\int_0^{N_Z} x_{Zj}^{\alpha} dj) Z^{1-\alpha}$. L and Z are assumed to be supplied inelastically and are industry-specific inputs. $x_h, h = L,Z$ are industry h intermediate goods. At a point in time, in industry h there are $N_h$varieties of intermediate good to serve as inputs in the production of the industry-specific final good. Recall that both blueprints and intermediate goods can be used exclusively in the industry in which they have been created. $Y_L$and Y_Z, are aggregated according to a CES technology with elasticity of substitution between the two commodities given by ε and ζ, $0 < \zeta < 1$ is a distribution parameter stating the relative importance of each commodity in the aggregation. The price of Y is taken as the numeraire, whereas P_L and P_Z are the prices for the two final goods. Therefore, it is possible to determine the price index for the final good sector.

As in the baseline model of horizontal innovation, each intermediate good is produced by a local monopolist which determines the size of the intermediate good it supplies, taking its demand function from final good profit maximization. If any intermediate good costs one unit of Y to be produced, then following the same steps outlined in the baseline model of horizontal innovation, we determine the optimal price for an intermediate good and the size of an intermediate good. Accordingly, we define the intermediate good producer's profits for each industry.

The instantaneous value for profits must be substituted inside the expression for the present-discounted value of profits, as this expression gives the value of a new firm. Value that must be compared to the cost for starting up a new firm to check the entry patterns in the economy. As usual, $V_h = \tilde{\pi}_h / r$, $h = L,Z$. The rate of returns is unique due to non-arbitrage in the asset market. The free entry assumption is at work, therefore the value of a new firm equals the R&D entry cost given by the price of a new patent.

Turning to R&D, it is assumed that the entry cost is an R&D cost determined by R&D firms' profit-maximization problem. The R&D cost that entrants must pay corresponds to the price of the new patent which gives room to new variety production. Each industry deals with its own R&D activity. However, partial excludability of technology allows for spillovers across industries.

To generate a final blueprint that can be sold in the market to entrant firms, the R&D firm i in industry h must undertake a costly activity described by the following technology $b_{ih} = \eta_h N_h^{\frac{1+\delta}{2}} N_k^{\frac{1-\delta}{2}} S_{hi}$, where $h = L,Z$ and $h \neq k$. b_{ih} corresponds to the single new blueprint produced by R&D firm i in industry h. S_{hi} are scientists employed by R&D firm i-th in the same industry, η_h is an exogenous productivity parameter. Scientists are supplied by households and the economy is endowed with a fixed amount of them, S, which is supplied inelastically to the two R&D activities. In the baseline model, there is a positive externality effect exerted by the knowledge stock. Here the positive effect is played both by knowledge generated in the same industry in which R&D takes place and by knowledge generated in the other industry. The degree of state dependence of a new blueprint from existing knowledge is measured by δ, $\delta \leqslant 1$: if $\delta = 1$, then inter-industry spillovers are ruled out; on the other extreme, if $\delta = 0$, then both intra- and inter-industry externality have the same effect.

Aggregating across each R&D sector, we identify the laws of motion for R&D:

$$\dot{N}_{ih} = \eta_h N_h^{\frac{1+\delta}{2}} N_k^{\frac{1-\delta}{2}} S_h$$

As in the baseline horizontal innovation set-up, solving for the profit maximization problem of the R&D firm i in sector h, and imposing that wages to scientists must be equalized across industries in equilibrium, it is possible to determine the R&D cost for entrants. The patent price is then used inside the free entry condition in each industry. Finally, assuming that non-arbitrage in the asset market holds in equilibrium, and applying other substitutions, we are able to identify the effects of R&D spillovers on equilibrium allocations. In fact, we end up with the following equation relating state variables to parameters only: $N_Z/N_L = (\eta_Z/\eta_L)^{\frac{\sigma}{1-\delta\sigma}}[(1-\zeta)/\zeta]^{\frac{\varepsilon}{1-\delta\sigma}}(L/Z)^{\frac{1-\sigma}{1-\delta\sigma}}$.

It is easy to notice that the stocks of knowledge exert two effects: one is channelled through the size of the intermediate good sector through prices of final goods. This effect does not depend on R&D spillovers and it will persist even if we assume that there are no R&D spillovers at all. The other effect is a direct consequence of R&D spillovers: it affects return from entry by lowering the entry costs in both industries. The influence of spillovers in shaping the direction of technological change goes through the relative technology bias, which in turns depends on: the relative factor supplies, L/Z; the elasticity of substitution between factors, ε; and R&D spillovers, summarized by the parameter δ. If the two factors are gross complements, $\sigma < 1$, then an increase in the relative factor supply decreases the relative technology bias. If the two factors are gross substitutes, $\sigma > 1$, then we have to distinguish between two scenarios: when the relative importance of inter-industry R&D spillovers is greater than intra-industry R&D spillovers, $\delta < 0$, an increase in the relative factor supply increases the relative technology bias; when intra-industry R&D spillovers are more important than the inter-industry one, then as long as $\delta > 1/\sigma$, an increase in the relative factor supply decreases the relative technology bias.

To check the influence of technology spillovers on growth, it is necessary to recall that we are assuming to be on the balanced growth path, therefore N_L and N_z must grow at the same rate, which also equals the growth rate of consumption given by the standard Euler equation. Therefore

$$\gamma = \eta_L \eta_Z S / \left[\eta_Z \left(\frac{N_Z}{N_L} \right)^{\frac{1-\delta}{2}} + \eta_L \left(\frac{N_L}{N_Z} \right)^{\frac{1-\delta}{2}} \right] \text{ with the ratio } N_Z / N_L \text{ determined above.}$$

R&D spillovers, growth and policy design A quick look to the reviewed papers is enough to notice the rich array of possible R&D spillovers which have been investigated in their relationship with growth.

Another noticeable result refers to the key role played by R&D activities and the associated externalities on equilibrium allocation, growth rate and policy design. R&D spillovers presented in the previous sections can be grouped in four groups: (1) positive intertemporal spillovers or the 'standing on the shoulders of giants' effect; (2) negative intertemporal spillovers as the most obvious ideas are discovered first; (3) the negative business-stealing effect; (4) consumers. Then, each group can exert its effects within R&D sector and/or throughout the whole economy, intra-industry or inter-industry.

Depending on the combination of R&D activities and spillovers considered in the model, we have seen that results change considerably, both with respect to the type of growth and with respect to optimal R&D policy. Due to the kind of distortion that they create in the economy, it is possible to generalize the optimal kind of fiscal tool for each type of R&D externality: (1) positive intertemporal spillovers call for public support; (2) negative intertemporal spillovers need taxation; as well as (3) the negative business-stealing effect; (4) consumers imply public aid. When a set-up deals with a unique source for R&D spillovers, or with more sources with the same direction, it is easier to assess the distance between the decentralized outcomes and the socially optimal ones. If spillovers are all positive, then the decentralized economy will underinvest in R&D as there will be a failure to internalize the positive externality effects. Therefore, the government must support private investments in R&D. Conversely, if spillovers are all negative, private agents will over-invest in R&D, since they are unable to identify the negative social side-effect of each private investment decision. Under this scenario, the government must tax private R&D investments.

However, if a set-up accounts for R&D spillovers with different signs, it allows for opposite forces influencing equilibrium allocations. In this case, determining whether the decentralized outcomes are either higher or lower than socially desirable becomes difficult. In fact it depends on the relative strength of opposite R&D spillovers. In this kind of framework, policy advice with respect to R&D becomes more complicated, as support may turn into taxation depending on the characteristics of the economy.

The wide set of growth predictions and policy advice coming from the literature on R&D and endogenous growth must not be interpreted as a hint for the absence of clear-cut results on the linkages between R&D and growth. It just states that, R&D being a complex object whose characteristics may vary depending on many variables, then, depending on the features it is actually clothed with, its effects on growth and welfare change dramatically. And endogenous growth literature can address a comprehensive set of possible features of R&D.

Table 12.1 Growth results in the surveyed models

Authors	Type of growth
Acemoglu (2002)	endogenous
Aghion and Howitt (1999)	endogenous
Jones (1995)	semi-endogenous
Li (2000, 2002)	semi-endogenous
Peretto (1999a, 1999b, 2003)	endogenous
Romer (1990)	endogenous
Segestrom (1998, 2000)	semi-endogenous
Strulik (2005)	endogenous

The way R&D is modelled has also a strong influence on the kind of growth experienced by the economy. We have seen that R&D-based models of growth may lead to either endogenous growth or semi-endogenous growth. The former is fully determined inside the economy, whereas the latter depends on the growth rate of the population, which is exogenous, although technical change is endogenous. Table 12.1 summarizes the type of growth associated with the surveyed models of endogenous technological change based on private R&D.

R&D spillovers and sustainable growth

In recent years there has been an increasing interest in extending endogenous growth models to incorporate environmental considerations, as the natural environment is acknowledged to be indirectly a factor of production. Allowing for environmental issues inside the economy leads to concerns about pollution, which is, to date, widely acknowledged to be an inevitable by-product of economic activity and to be directly and negatively related to the level of environmental quality (Bovenberg and Smulders, 1996).

Many contributions have explored frameworks where innovation plays a key role in abating pollution (Bovenberg and Smulders, 1996; Smulders and Gradus, 1996). As the aim of the chapter is not to analyse the linkage between R&D and the environment, we simply provide some short considerations about the topic, mainly referring to policy design for sustainable growth. Generally, when pollution is considered along with innovation, the focus is limited to R&D explicitly pursued to abate pollution. This perspective follows from the fact that R&D being carried on by firms for profit motives, it aims at new product discoveries or cost reduction and, as long as environmental protection does not generate profits, firms must be pushed towards pollution-abating R&D investment through government deeds.

Basic research is quite likely to be the best candidate to exert positive consequences on pollution abatement. It is noteworthy that the US Environmental Protection Agency is currently monitoring R&D in nanofields (also privately performed), although not environmentally oriented, to understand and evaluate the positive environmental side-effects (EPA, 2005). It has been estimated that the environmental trajectories of nanotechnology, chemical and manufacturing R&D lead to potential energy savings for the US close to 14.5 per cent of total US energy consumption per year (EPA, 2005). Not one of these technologies has been developed for environmental protection and they rely heavily on

basic research. There is also anecdotal evidence about unintended pollution-reducing trajectories of privately performed basic research-intensive activities. Then, several works about nanotechnology roadmapping have highlighted important environmental side-effects of nanotech knowledge: from improved healthcare[9] to environmental protection and energy savings (EPA, 2005; Ernst and Shetty, 2005). This newer perspective may also give some appealing insights with respect to the linkage between R&D and environmental policy. In the environmental literature, R&D policy is generally needed to push private firms to perform the socially optimal amount of pollution-abating R&D. However, evidence on federal support to R&D in the US shows that different fiscal incentives are used depending on R&D composition, and that federal support is mainly directed towards basic research activity. Therefore, by constructing a set-up where basic research is disentangled from development and contributes accidentally to environmental preservation, it is possible to check whether the observed differences in fiscal incentives also have some effects on environmental protection.

The consequences of this exercise may be particularly interesting considering that, notwithstanding well-established results on the damage created by pollution, environmental policy is enforced quite slowly, since efforts towards environmental improvements may be overshadowed by the fear that environmental policy damages the economy through a reduction in production and in economic growth. In fact, there is an active debate between those who argue that, pollution being an inevitable side-product of production, economic growth cannot be ecologically sustainable, and those who maintain that a growing economy can produce a growing amount of abatement devices so that pollution is offset (Smulders and Gradus, 1996). Empirical evidence for developed countries seems to support the latter point of view, as there exists substantial evidence that developed countries experience economic growth associated to improvement in environmental quality and that this is achieved through policy enforcement (Grossman and Krueger, 1995; Stokey, 1998).

So, if environmental policy coincides with R&D policy and R&D policy helps firms in their production-oriented R&D investments, then it may be that support to pollution abatement is enforced through support to privately performed and growth-promoting R&D. In this way we make a point in favour of the optimistic point of view about sustainable growth.

12.4 R&D externality and growth

Surveying theoretical contributions to the relationship between R&D spillovers and growth, we have found a wide array of results, driven by the different aspects of R&D considered. Multiplicity of results determines variety of policy tools. Theory on R&D as a commodity stresses that the externality effects exerted by technology vary depending on some features: structure of the R&D process, type of copyright law, components of the R&D process, and so on. Surveying some theoretical contributions on the relationship between R&D and economic growth, we have seen that results are hugely affected by these features. From a theoretical perspective there are no universal results, meaning that also policy design varies dramatically.

To shed some light on this, econometric analysis of the effects of R&D on growth play a pivotal role. However, adding data analysis does not allow us to identify a standard universal set of findings. Heterogeneity is due to a wide range of factors: different method-

ologies, major structural differences in both countries and time periods covered, and choices about the specification for R&D.

Assessing R&D spillovers

A considerable bulk of empirical studies has acknowledged that social rates of return from R&D are higher than private rates of return (for example Aiello and Cardamone, 2005; Cincera, 2005; Griliches and Lichtenberg, 1984). This evidence supports the prevalence of positive spillovers from R&D that cannot be internalized by firms and investors, as advocated from the baseline models of endogenous technological change. Private rate of return is generally measured through the impact that a given R&D investment by firm *i* exerts on its final output. Social returns are measured through the effect on firm *i* final output after R&D investment performed by other firms.

The standard model to measure the effects of R&D on GDP hinges on endogenous growth literature. Output is produced through a Cobb–Douglas technology $Q_{it} = A^{\lambda t} L_{it}^{\beta} K_{it}^{\alpha} R_{it}^{\gamma} X_{it}^{\eta} e^{\varepsilon_{it}}$, where Q is output of firm i at time t, A is an exogenous parameter, L is labour employed in final good production, K is capital input, R is the R&D input and X is R&D spillovers. e accounts for unknown factors and noises. Returns from R&D are determined considering the elasticity of GDP with respect to R&D direct input: $\gamma = (\partial Q_{it}/\partial R_{it})[Q_{it}/R_{it}]$, where $\partial Q_{it}/\partial R_{it}$ is the rate of return of the R&D input. Then, transformation of the expression for Q gives several models to test through data. The most common are the logarithmic and the first-difference transformations $q_{it} = \alpha\lambda t + \beta l_{it} + \alpha k_{it} + \gamma r_{it} + \eta x_{it} + e\varepsilon_{it}$ and $\Delta q_{it} = \alpha\lambda + \beta\Delta l_{it} + \alpha\Delta k_{it} + \gamma\Delta r_{it} + \eta\Delta x_{it} + e\Delta\varepsilon_{it}$. Both these specifications – and many others – suffer from some methodological problems: the way R&D stock is defined and measured, double counting with respect to other input accountancy, hedonic price deflation and spillovers.[10]

With respect to the latter, typically R&D spillovers are measured through the sum of R&D stocks of other firms, weighted by some index

$$X_{it} = \sum_{i \neq j} \omega_{ijt} R\,\&\,D_{jt},$$

where X_{it} is the R&D externality, $R\,\&\,D_{jt}$ is the direct R&D investment by firm j and ω_{ijt} is the weight shaping the way R&D effort by firm j affects firm i. ω_{ijt} may represent different features: geographical distance, technological distance, absorptive capacity, and so on. Many works have analysed the effects of R&D spillovers and they generally provide support for the beneficial effects exerted by R&D externality on the economy (for example Wieser, 2005). Generally, these works deal with technological spillovers.

Some authors argue that the estimated gap between social and private returns from R&D may be underestimated, as the standard approach neglects important channels through which other positive spillovers might play a role: R&D improves the ability of both imitation and adoption of existing technology; the spatial location of firms influences the relative strength of spillovers. Below we discuss some recent works exploring these issues and providing support to this point.

An important element supporting the influence of R&D and R&D spillovers on economic performance comes from the assessment of the direction of the causality relationship between the three variables. Endogenous growth literature based on R&D assumes that innovation causes growth through productivity improvements, a

hypothesis that needs to be tested. Recently, Lu et al. (2006) have conducted an empirical analysis of the causality relationship between R&D and productivity growth, accounting also for R&D spillovers. The analysis has been performed applying a Granger causality test to a sample of 90 firms for a time period of ten years. The results confirm that the causality goes from R&D and R&D spillovers to total factor productivity. Therefore, there is an empirical support for the key assumption beyond R&D-based endogenous growth.

Space

Important issues in need of empirical assessment refer to international spillovers from R&D and to the relationship between spillovers and the geographical distance of countries and regions.

In modelling these topics, there are mainly two approaches: some authors attach the same weight to each R&D contribution, no matter what firm it comes from, whereas other authors set a weight system considering each firm's location in a determined space (patent space, geographical space, and so on). The latter vision is supported by some relevant analysis acknowledging the role of agent interactions, non-market environment and proximity in favouring R&D diffusion (Audretsch and Feldman, 2004).

At the micro level, it is possible to establish the spatial dimension of R&D influence by tracking the pathways of knowledge diffusion. One way to determine these paths consists of submitting surveys to firms asking them about the impact of R&D performed by other firms, universities, and so on, on their productivity. The collected data allow the existence of technology spillovers and the importance of spatial distance in shaping the appropriability of spillovers to be determined (Doring and Schellenbach, 2006). Alternatively, patent citations on the same technology by firms and universities that are spatially close and that have happened really close in time constitute an index of the importance of spatial location in shaping R&D spillovers.

At a more aggregated level, the role of location in R&D and R&D spillovers is analysed starting from the following model:

$$Z_{si} = \alpha R^{\beta_1} (UR)_{si}^{\beta_2} [(GC)_{si}^{\beta_3} * (UR)_{si}] e_{si}$$

where s is a space index and i an industry index. Z is R&D output, R is private corporate R&D investment, UR is public R&D and GC measures the geographic coincidence between public and private R&D performers, and e accounts for unknown factors and noises (Audretsch and Feldman, 2004). Estimation of this model and of other models that can be tracked back to this baseline specification have shown that technology, the output of R&D, depends on space, through innovative efforts taking place nearby. Similar results can be found also using the patent citation methodology (Audretsch and Feldman, 2004).

To determine the role of foreign R&D spillovers on domestic growth, many authors have developed econometric models linking domestic total factor productivity (TFP) to both domestic and foreign R&D. The typical model is formalized as follows:

$$\ln TFP_{Ht} = \beta_H + \beta_t + \lambda_r \ln(R \,\&\, D)_{Ht} + \lambda_s \ln(R \,\&\, D_F)_{Ht} + u_{Ht}$$

where H stands for the domestic country and F for foreign. u_{Ht} is an error term. Foreign R&D is given by a weighted sum of R&D performed in other countries,

$$R\,\&\,D_F = \sum_{i \neq H} \omega_{Hit} R\,\&\,D_{it},$$

where the weights measure the degree of interdependence between countries. Interdependence can be input–output shares, import shares, foreign direct investment, and so on (Gong and Keller, 2003). In many contributions, international R&D spillovers are found to be significant in positively influencing domestic TFP. Therefore, there is evidence supporting a positive effect of foreign R&D on domestic growth. Having assessed that R&D exerts positive spillover effects under some circumstances, it is interesting to identify the pervasiveness of these unintended positive effects – in particular, the link between country distance and spillover strength.

Keller (2002) considers the effect on domestic TFP of R&D efforts done in the five most advanced economies, allowing for some influence of the distance between the domestic country and the top five economies:

$$\ln TFP_{Hit} = \lambda \ln\left(R\,\&\,D_{Hit} + \gamma \sum_{F \neq H} (R\,\&\,D)_{Fit} e^{-\delta D_{HF}} \right) + \underline{\beta'\, X} + u_{Hit}$$

The spatial distance between the domestic country and the foreign one is measured by the term D_{HF}. Estimation of the model using industry-level data from 14 Organisation for Economic Co-operation and Development (OECD) countries for the period 1970–95 gives a positive value for δ, implying that distance matters in shaping the influence of foreign R&D. Moreover, there is evidence that the distance parameter becomes smaller as time passes. Therefore, distance matters, but at the same time knowledge is becoming less country-specific.

To address this question, Bottazzi and Peri (2003) test for the existence of localized spillovers from R&D by checking the effects of R&D investments in region a on productivity of R&D in region b. Formally, the R&D production function for region i is given by:

$$\Delta A_i = B(R\,\&\,D)_i^{e_R} A_i^{e_0} \prod_{i \neq j} A_j^{e^{(dist_{ij})}},$$

where ΔA_i measures the change in the stock of knowledge for the period under consideration. Knowledge creation depends on R&D resources, *R&D*, and on the existing stock of knowledge, measured by A_i, for the knowledge generated in region i, and A_j for the knowledge generated in region j. B is a constant term accounting for the common factors and e_k measures the elasticity of innovation of each input: e_R measures the elasticity of innovation to R&D input, e_0 is the elasticity of the existing stock of knowledge generated within the region, $e_{(\text{dist}_{ij})}$ measures elasticity of the existing stock of knowledge generated in region j. The latter depends on the distance between the region of production, j, and the region under study, i. It is assumed that there is a maximum distance, K, beyond which knowledge does not spill. Assuming the balanced growth path, through log-linearization of the baseline equation it is possible to rewrite the expression in the following way:

$$\ln(\Delta A)_i = \beta + \varepsilon_0 \ln(R\&D)_i + \varepsilon_{[dist0dist1]}[\underline{m'}_{i1}\ln(R\&D)] + \ldots + \varepsilon_{[distndistK]}[\underline{m'}_{iK}\ln(R\&D)] + u_i$$

where $\varepsilon_0 = e_R/(1-e_0)$, $\varepsilon_k = e_R e_k/(1-e_0)^2$. The model is estimated using a cross-section of long-run averages of variables referring to 86 regions for the period 1977–95. ΔA_i is measured through average yearly patent applications in region *i* filed with the European Patent Office, $R\,\&\,D_i$ by the average yearly real spending for R&D in region *i*. $\underline{m'}_{iK}\ln(R\,\&\,D)$ stands for the average *lnR&D* for each region at distance *k* from region *i*. Finally, u_i is independently identical distributed error. The authors consider five distance classes among regions: 0–300, 300–600, 600–900, 900–1300, 1300–2000 km. Then, they add a dummy for unobserved factors.

Ordinary least squares (OLS) estimation implies that local spillovers exist only within 300 km and than interregion R&D spillovers are weaker than intra-region R&D spillovers. These findings are robust even if we control for industry composition in each region and for human capital.

Recently, Rodríguez-Pose and Crescenzi (2006) have developed an econometric model where GDP growth is assumed to be dependent on R&D investments, R&D spillovers and socio-economic conditions in both the domestic economy and the neighbouring regions. The novelty resides in having embedded social aspects as means to capture the influence of local aspects on the diffusion and the effects of R&D spillovers. This approach accounts for the claim that knowledge is sticky and it propagates better if there are proper and frequent interactions between economic agents. Formally, the model is represented by the following equation:

$$(1/J)\ln(Y_{it}/Y_{it-J}) = \alpha + \beta_1 \ln(y_{i,t-J}) + \beta_2 R\,\&\,D_{it-J} + \beta_3 SocFilter_{it-J} + \beta_4 Spillo_{it-J} + \beta_5 ExtSocFilter_{it-J} + \beta_6 ExtGDPcap_{it-J} + \beta_7 D + \varepsilon$$

where on the left-hand side there is an expression for the log of the ratio of regional GDP per capita between the two extremes for the time period. On the right-hand side, besides the constant α and the error term ε, we deal with log of regional GDP per capita, regional R&D investments, a proxy for socio-economic regional characteristics, a proxy for the region's ability to access non-domestic innovation, a proxy for neighbouring socio-economic characteristics, GDP per capita in neighbour regions and a dummy for national effects. Social filters are measured by means of a composite index containing educational, demographic and employment aspects.

The model is tested on the 25 European Union member countries for the period 1955–2003. The methodology used hinges on OLS. The results show that R&D spillovers are able to exert positive effects only if the recipient region is endowed with social characteristics such that R&D flows are understood and adopted.

Convergence and R&D spillovers

R&D spillovers have a key role also in determining the path of convergence among regions and countries. Patterns on increasing integration among countries, as the steps undertaken by the European Union since the 1980s, determines a background that favours knowledge spillovers as cultural and structural boundaries are reduced and ideas can circulate more easily. In turn, R&D externalities affect convergence across the member countries.

Lately, Giannetti (2002) has addressed this issue by constructing a model able to capture the influence of increased integration on neighbouring regions on spillovers and convergence. The theoretical predictions of the model suggest that: (1) increasing interaction among countries increases knowledge spillovers; moreover, (2) knowledge spillovers have different effects depending on the technological compositions of regions in each country: for high-tech regions, the increase in R&D spillovers exerts positive effects towards convergence, whereas for low-tech regions the opposite applies. Overall, the effects of increased integration on R&D spillovers are such that there is a push towards polarization between high-tech regions and low-tech regions.

Predictions are empirically tested through a growth regression for EU countries for the years 1980–92. Growth rate of GDP depends on productivity of regions, considered both at the aggregate level and weighted both in terms of technological level (high- versus low-tech region) and in terms of sub-periods. Empirical results support the theoretical claim that country convergence does not match with regional convergence. Indeed, with respect to regions, there is evidence of increased polarization due to differences in exploiting knowledge spillovers.

R&D components

Another interesting topic refers to the spillovers effects associated to the different components of an R&D process. From a theoretical point of view, we have seen that there are meaningful distinctions among R&D activities, and between R&D and learning-by-doing. Some empirical works have assessed the relevance of some of these distinctions and the significance of the associated spillovers.

We have seen above that disentangling basic research from applied research and development allows their different economic features to be accounted for; in particular, basic research being more general and with a higher innovative content, it should be associated with larger and more pervasive spillovers. A recent analysis on international research spillovers exerted by R&D components on a panel of nine OECD countries for the period 1981–93 confirms the theoretical predictions: basic research generates larger international spillovers than developmental research (Funk, 2002). Using dynamic OLS methodology, it is possible to consider the following model:

$$TFP_{Ht} = \alpha_i + \beta_{DHt} D_{Ht} + \beta_{DFt} D_{Ft} + \beta_{BFt} B_{Ft} + \sum_{j=-q^1}^{q^2} c_{DHj} \Delta D_{Ht+j}$$

$$+ \sum_{j=-q^1}^{q^2} c_{DFj} \Delta D_{Ft+j} + \sum_{j=-q^1}^{q^2} c_{BFj} \Delta B_{FHt+j} + u_{Ht}$$

where D_{Ht} stands for domestic expenditure on development, D_{Ft} for foreign expenditure on development and B_{Ft} for foreign expenditure on basic research. Then, the dynamic OLS technique allows serial correlation and endogeneity to be controlled for by adding lags and leads of differenced regressors. The estimation results reveal that domestic development spills only internally, although it exerts an important contribution on domestic TFP. Basic research generates substantial international spillovers. For the nine OECD countries considered, foreign basic research is nearly as important to productivity growth as its own development.

Another distinction considers the roles of R&D: R&D serves both to promote innovation and to facilitate imitation. As a consequence, R&D spillovers also happen through improvements in the ability to understand innovations made by others, and therefore they improve the ability to imitate (absorptive capacity). Moreover, it is also possible to account for the effect of R&D on the relative position of the country with respect to the frontier of knowledge (technology transfer). These features have recently been explored by Griffith et al. (2004) for a panel of 12 OECD countries since 1970. The baseline model tests the dependency of TFP on R&D, absorptive capacity and technology transfer:

$$\ln(\Delta A)_{ijt} = \rho(R/Y)_{ijt-1} + \beta_1 \ln(A_F/A_i)_{jt-1} + \beta_2 (R/Y)_{ijt-1} \ln(A_F/A_i)_{jt-1} + \gamma X_{ijt-1} + u_{ijt}$$

where the first term captures the elasticity of GDP with respect to the R&D stock, the second captures the technology transfer, A_F being the stock of knowledge of the technologically leading country, and the third stands for absorptive capacity. It is straightforward to notice that the absorptive capacity effect is a combination of the other effects. Moreover, for the technologically leading country, the second and the third terms disappear. Then, X_{ijt-1} is a vector of control variables and u_{ijt} errors. Index i labels countries, j industries and t time. The application of the model to the panel shows that both effects are empirically verified: R&D benefits TFP both through improving the ability to understand innovations in general and also through technology transfer from leading countries.

12.5 Conclusions

Spillovers are one of the main features characterizing R&D as a commodity. Their existence is both identified by the theory and supported by the data. Moreover, both theoretical and applied literature on R&D have highlighted a wide menu of trajectories through which R&D externalities may influence an economy: some of them have positive consequences, other negative. Obviously, spillovers benefiting an economy must be supported by policy design, whereas negative influences have to be reduced. However, R&D is not an easy object to deal with: the literature shows clearly that depending on the combination of R&D activities under consideration, the effects on the economy change dramatically: from policy design to the type of growth the economy experiences.

Empirical evidence cannot provide a unique clear-cut result about the effect of R&D spillovers, either. Nonetheless, it provides support to some of the most important findings determined by the theorethical literature. First of all, there is support for the causal relationship between R&D and growth, the assumption guiding all the bulk of literature on endogenous technological change. Then, many works provide support for the beneficial effects exerted by R&D externality on the economy, even though there are space constraints and pervasiveness changes depending on the type of R&D activity under consideration.

Notes

1. US National Science Foundation.
2. This strategy is called 'Google 20 percent'. Some of Google's newer services and products have originated from these independent endeavours (Mayer, 2006).
3. 'Although risk is associated with all forms of R&D, uncertainty is an inherent characteristic of basic research. Not surprising that the outcome and the direction of basic research is often unpredictable'; 'Moving from the applied research end of the spectrum to the basic research end, the degree of uncertainty about results of specific research projects increases and the goals become less clearly defined and less closely linked to the solution of a specific object' (Nelson, 1959).

4. There are many examples of private firms declaring to invest in basic research for strategic reasons, for example Microsoft Corp., Intel Corp., IBM, Bell Labs, Google.
5. In particular, we abstract from capital accumulation. This assumption does not influence the main results. We will also stick to this simplification in the successive sections of the chapter.
6. This is a simplified version of the baseline model discussed in Aghion and Howitt (1999).
7. If $\phi = 1$ and $\lambda = 1$, the model reduces to the baseline model described in the previous section.
8. Li (2002) states that the q in the denominator is used for expositional purposes and it does not constitute a key element in driving the results.
9. It is widely acknowledged that pollution has a positive and significant effect on cancer proliferation and on reducing the quality of life (Arden et al., 2002)
10. For a detailed discussion on this point, see Wieser (2005).

References

Acemoglu, D. (2002), 'Directed technical change', *Review of Economic Studies*, **69**, 781–809.

Acemoglu, D. and F. Zilibotti (2001), 'Productivity differences', *Quarterly Journal of Economics*, **116**, 563–606.

Aghion, P. and P. Howitt (1999), *Endogenous Growth Theory*, Cambridge, MA: MIT University Press.

Aiello, F. and P. Cardamone (2005), 'R&D spillovers and productivity growth: evidence from Italian manufacturing microdata', *Applied Economics*, **12**, 625–31.

Arden, P., R.T. Burnett, M.J. Thun, E.E. Calle, D. Krewski, K. Ito and G.D. Thurston (2002), 'Lung cancer, cardiopulmunary mortality and long term exposure to fine particulate air pollution', *Journal of American Medical Association*, **287**, 1142–6.

Arrow, K.J. (1962), 'Economic welfare and the allocation of resources for inventions', in R.R. Nelson (ed.), *The Rate and Direction of Inventive Activity*, Princeton, NJ: Princeton University Press for NBER, pp. 609–26.

Audretsch, D.B. and M.P. Feldman (2004), 'Knowledge spillovers and the geography of innovation', in V. Henderson and J.F. Thisse (eds), *Handbook of Regional and Urban Economics*, **4**, Amsterdam: Elsevier, pp. 2713–39.

Audretsch, D.B., B. Bozeman, K.L. Combs, M. Feldman, A. Link, D. Siegel, P. Stephan, G. Tassey and C. Wessner (2002), 'The economics of science and technology', *Journal of Technology Transfer*, **27**, 155–203.

Auerswald, P.E., L.M. Branscomb, N. Demos and B.K. Min (2005), 'Understanding private-sector decision making for early-stage technology development', NIST Advanced Technology Program publication series NIST GCR02-841A.

Barro, R., and X. Sala-i-Martin (2005), *Economic Growth*, Cambridge, MA: MIT University Press.

Bottazzi, L. and G. Peri (2003), 'Innovation and spillovers in regions: evidence from European patent data', *European Economic Review*, **47**, 687–710.

Bovenberg, A.L. and S. Smulders (1996), 'Transitional impacts of environmental policy in an endogenous growth model', *International Economic Review*, **37**, 861–93.

Branscomb, L.M. and P.E. Auerswald (2001), *Taking Technical Risks*, Cambridge, MA: MIT University Press.

Cincera, M. (2005), 'Firms' productivity growth and R&D spillovers: an analysis of alternative technological proximity measures', *Economics of Innovation and New Technology*, **14**, 657–82.

David, P.A. (1997), 'From market magic to calypso science policy: a review of Terence Kealey's *The Economic Laws of Scientific Research*', *Research Policy*, **26**, 229–55.

Doring, T. and J. Schellenbach (2006), 'What do we know about geographical knowledge spillovers and regional growth? A survey of the literature', *Regional Studies*, **40**, 375–95.

Eisenman, E., K. Koizumi and D. Fossum (2002), 'Federal investment in R&D', RAND publication MR-1639.0-OSTP, http://www.rand.org.

Environmental Protection Agency (EPA) of the United States (2005), 'Nanotechnology White Paper', www.epa.gov.

Ernst, H. and R. Shetty (2005), 'Impact of nanotechnology on biomedicalsciences: review of current concepts on convergence of nanotechnology with biology', *Journal of Nanotechnology Online*.

European Commission, Community Research (2002), 'Fusion energy: moving forward', EUR 20229.

Funk, M. (2002), 'Basic research and international spillovers', *International Review of Applied Economics*, **16**, 217–26.

Gancia, G. and F. Zilibotti (2005), 'Horizontal innovation in the theory of growth and development', in P. Aghion and S. Durlauf (eds), *Handbook of Economic Growth*, Vol. 1, Amsterdam: Elsevier, pp. 111–70.

Giannetti, M. (2002), 'The effects of integration on regional disparities: convergence, divergence or both?', *European Economic Review*, **46**, 539–67.

Gong, G. and W. Keller (2003), 'Convergence and polarization in global income levels: a review of recent results on the role of international technology diffusion', *Research Policy*, **32**, 1055–79.

Griffith, R., S. Redding and J. Van Reenan (2004), 'Mapping the two faces of R&D productivity growth in a panel of OECD industries', *Journal of International Trade and Economic Development*, **17**, 105–33.

Griliches, Z. and F. Lichtenberg (1984), 'Interindustry technology flows and productivity growth: a re-examination', *Review of Economics and Statistics*, **86**, 883–95.
Grossman, G.M. and E. Helpman (1989), 'Product development and international trade', *Journal of Political Economy*, **97**, 1261–83.
Grossman, G.M. and E. Helpman (1991), 'Innovation and growth in the global economy', Cambridge, MA: MIT Press.
Grossman, G.M. and P. Krueger (1995), 'Economic growth and the environment', *Quarterly Journal of Economics*, **110**, 353–77.
Jacobs, B., R. Nahius and P.J.G. Tang (2002), 'Sectoral productivity growth and R&D spillovers in the Netherlands', *De Economist*, **150**, 181–210.
Jones, C.I. (1995), 'R&D-based models of endogenous growth', *Journal of Political Economy*, **103**, 759–84.
Jones, C.I. (1999), 'Growth: with or without the scale effect?', *American Economic Review*, **89**, 139–44.
Judd, K.L. (1985), 'On the performance of patents', Econometrica, **53**, 567–85.
Kealy, T. and A. Rudenski (1998), 'Endogenous growth theory for natural scientists', *Nature Medicine*, **4**, 995–9.
Kesteloot, K. and R.Veugelers (1995), 'Stable R&D cooperation with spillovers', *Journal of Economics and Management Strategy*, **4**, 651–72.
Li, C. (2000), 'Endogenous vs. semi-endogenous growth in a two-R&D-sector model', *Economic Journal*, **110**, C109–C122.
Li, C. (2002), 'Growth and the scale effects: the role of knowledge spillovers', *Economics Letters*, **74**, 177–85.
Lichtenberg, F.R. and D. Siegel (1991), 'The impact of R&D investment on productivity: new evidence using linked R&D–LRD data', *Economic Inquiry*, **29**, 203–29.
Lu, W.C, J.R. Chen and C.L. Wang (2006), 'Granger causality test on R&D spatial spillovers and productivity growth', *Applied Economics Letters*, **13**, 857–61.
Mayer, M. (2006), 'Speeches on MS&E entrepreneurial thoughts', ETL Seminar Series, Stanford University.
National Science Foundation (NSF) (2004), 'National patterns of research and development resources: 2003', available at http://www.nsf.org.
Nelson, R.R. (1959), 'The simple economics of basic scientific research', *Journal of Political Economy*, **67**, 297–306.
Pavitt, K. (2001), 'Public policies to support basic research: what can the rest of the world learn from US theory and practice? (and what they should not learn)', *Industrial and Corporate Change*, **10**, 761–79.
Peretto, P. (1999a), 'Cost reduction, entry, and the interdependence of market structure and economic growth', *Journal of Monetary Economics*, **43**, 173–95.
Peretto, P. (1999b), 'Firm size, rivalry and the extent of the market in endogenous technological change', *European Economic Review*, **43**, 1747–73.
Peretto, P. (2003), 'Fiscal policy and long-run growth in R&D-based models with endogenous market structure', *Journal of Economic Growth*, **8**, 325–47.
Rodríguez-Pose, A. and R. Crescenzi (2006), 'R&D, spillovers, innovation systems and the genesis of regional growth', Bruges European Economic Research Papers, N.5.
Romer, P. (1990), 'Endogenous technological change', *Journal of Political Economy*, **98**, 71–102.
Segestrom, P.S. (1998), 'Endogenous growth without the scale effects', *American Economic Review*, **88**, 1290–1310.
Segestrom, P.S. (2000), 'The long-run growth effects of R&D subsidies', *Journal of Economic Growth*, **5**, 277–305.
Smulders, S. and R. Gradus (1996), 'Pollution abatement and long-term growth', *European Journal of Political Economy*, **2**, 505–31.
Stokes, D. (1997), *Pasteur's Quadrants: Basic Science and Technological Innovation*, Washington, DC: Brookings Institution Press.
Stokey, N. (1998) 'Are there limits to growth?', *International Economic Review*, **39**, 1–31.
Strulik, H. (2005), 'The role of human capital and population growth in R&D-based models of economic growth', *Review of International Economics*, **13**, 129–45.
Theis, T.N. and P.M. Horn (2003), 'Basic research in the information technology industry', *Physics Today*, **57**, 44–52.
United States Office for Management Budget (2004), 'Fiscal year 2004 Budget report', www.whitehouse.gov/omb/.
Wieser, R. (2005), 'Research and development productivity and spillovers: empirical evidence at the firm level', *Journal of Economic Surveys*, **19**, 587–621.

13 Knowledge and regional development

Börje Johansson and Charlie Karlsson

13.1 Introduction

This chapter examines models depicting and explaining the role of knowledge in regional development and provides an assessment of empirical studies of how knowledge affects growth and development in functional regions. In this endeavour, it is crucial to understand those factors that make knowledge spatially sticky and knowledge-production capacity trapped. It is equally important to explain the conditions for knowledge flows and diffusion. The presentation also widens the view by linking knowledge generation to creativity.

Regions and regional development

In recent decades, the world has witnessed the emergence of a global knowledge economy, in which regions are increasingly being looked upon as independent, dynamic market places, which are connected with each other via knowledge and commodity flows. Each such region has its own base of scientific, technological and entrepreneurial knowledge, framing the conditions for regional growth.

The above picture provides a meaningful description only when we have a clear concept of what a region is. We shall use the concept of 'functional' (urban) region, as a place for knowledge creation, appropriation and absorption, as well as a place for transforming knowledge to innovations (Jaffe et al., 1993; Glaeser, 1999; Karlsson and Andersson, forthcoming). In essence, a region is an arena for exploiting communication externalities by means of face-to-face interactions (Fujita and Thisse, 2002), Another associated aspect is the concept of labour market region, in which knowledge spreads as individuals change their job affiliation (Zucker et al., 1998).

Localised knowledge

Although we have not yet provided any categorisation of knowledge, we may ask a preliminary question: which types of knowledge tend to be clustered spatially? Knowledge in the form of firm assets, such as patents, employed technology and research and development (R&D) capacity will be concentrated in space to the extent that knowledge-rich firms are co-located in the same region. Knowledge in the form of human capital becomes localised as the result of a clustering process, where concentrations of persons embodying knowledge and creativity attract knowledge-intensive individuals to migrate to such places and to remain there.

Localised knowledge will have a sustainable influence on a region's future development if the knowledge resources of the region change on a slow time scale. When this applies, a region with small knowledge resources can accumulate more knowledge only over an extended time period, whereas a knowledge-rich region will tend to remain such far into the future.

History offers examples where household migration and firm relocation have been fast, but slow adjustments remain the rule, and large urban regions often have a history

measured in centuries. Two basic factors explain this temporal phenomenon. First, infrastructure and associated amenities operate as slowly changing attractors for both firms and households. Second, with reference to new economic geography, the following cumulative location externality applies:

1. Knowledge-intensive labour is attracted to places where knowledge-dependent firms are located.
2. Firms with knowledge-dependent activities are attracted to places where knowledge-intensive labour resides.

13.2 The nature of knowledge in the economy

The concept of knowledge is elusive and interpretations easily become deceitful. Therefore, we will discuss the nature of knowledge and its effects on the economy. We also ask: from where does knowledge come and why is it sticky?

The form of knowledge

Authors like Kobayashi (1995) have been careful to distinguish between information and knowledge. From one point of view, information is simply a carrier of messages, and such messages can contain statements about knowledge. It is perhaps in this context that the distinction between information and knowledge is most vital, because some forms of knowledge are difficult (or impossible) to codify and thus to transform into useful messages.

Focusing on knowledge that can be related to production activities of firms and other organisations (like public authorities, and so on), we recognise the following three categories:

- Know-how, which is always embodied in persons or embedded in an organisation.
- Know-why, which has the nature of systematic and publicly accepted (scientific) explanations, which can be stored in codified form, but may require skill to decode.
- Knowledge in the form of human capital, which represents both know-how and know-why, embodied in individuals.

Know-why refers to a capacity to explain and understand, whereas know-how signifies expertise, skills and practical attainments. An entrepreneur or a salesperson may be skilful in finding customers and making them buy the products offered for sale, while at the same time being unaware of why the marketing methods are successful.

Know-why relates to science in the sense that it does not exist – by definition – if it has not been codified. In contradistinction, know-how can be present without codified instructions, generically based on experience and training and often so difficult (or uneconomical) to codify that it becomes tacit.

From another perspective, Karlsson and Johansson (2006) introduce the following three knowledge concepts: scientific (principles), technological (blueprints) and entrepreneurial (business) knowledge. In this context, it seems important to remark that both scientists and engineers perform R&D activities, while making use of know-how about effective and feasible ways to conduct research. Attempts to codify such know-how are often quite primitive and superficial.

For innovations, it is essential to consider the degree to which knowledge is 'rivalrous' and 'excludable' (cf. Cornes and Sandler, 1986). A rival good has the property that its use by one actor precludes its use by another, whereas a non-rival good lacks this property. Excludability relates to both technology and legal systems (Kobayashi and Andersson, 1994). A good is excludable if the owner can prevent others from using it. Pure public goods are both non-rival and non-excludable. As stressed by Arrow (1962), this creates a conflict, since a firm will only be motivated to carry out R&D if competitors can be excluded, whereas society will benefit if the knowledge (innovation) is allowed to diffuse to many firms.

How does the knowledge affect the economy?
What types of knowledge can be identified for a firm? The following three components comprise a firm's primary types of knowledge:

- knowledge about firm routines;
- knowledge about product varieties;
- knowledge as a capacity to carry out R&D.

Firm routines include techniques and approaches that are applied in production, administration, logistics, distribution and transaction activities. In this way, the routines (production technique) are a manifestation of the firm's know-how, where the latter also includes the firm's capability to combine product attributes of its output varieties. For an innovative firm, routines may also comprise its procedures to improve – gradually or stepwise – its routines and to develop its product varieties (Nelson and Winter, 1982).

Given the structure outlined in Figure 13.1, how do economic models describe the influence of knowledge? When answering this, we identify two approaches in orthodox theory:

- Knowledge affects (augments) the production function of a firm, which implies that it improves the productivity of inputs (Chambers, 1988).
- Knowledge affects the value ladder of product varieties produced by the firm (Grossman and Helpman, 1991).

The microeconomic production function isolates the study of knowledge to a question of improving routines of established and potential firms. However, it excludes an important temporal phenomenon by not considering the routines a firm has available when improving its production process. When such issues are at the forefront, researchers take the step into industrial dynamics and evolutionary economics (Nelson and Winter, 1982; Dosi et al., 1988).

How can we take the step from Figure 13.1 to a model of how knowledge affects the economy of a region? Following a mainstream approach leads to a regional production-function model of the following kind:

$$Y = F(K, L, N)A \tag{13.1a}$$

$$A = \int_{t} \dot{A} dt \tag{13.1b}$$

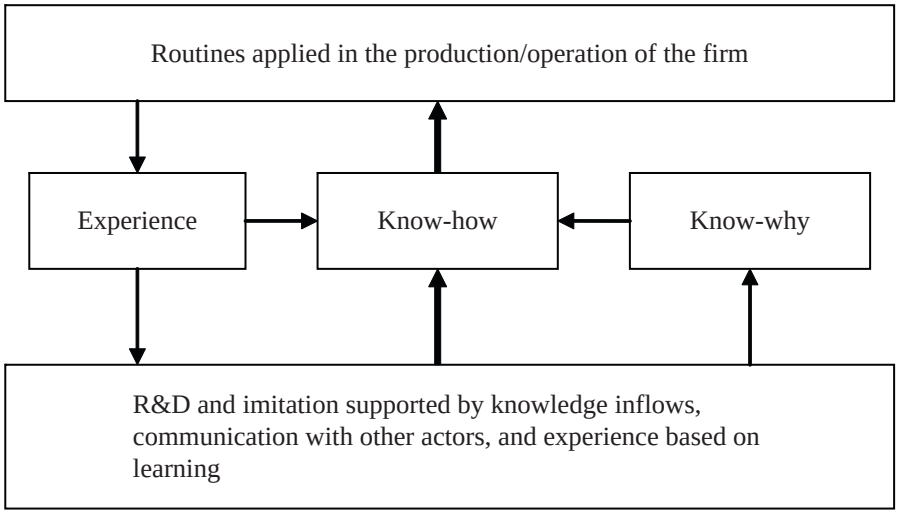


Figure 13.1 Knowledge development of a firm

$$\dot{A} = G(\hat{K}, \hat{L}, \hat{N})A \tag{13.1c}$$

where K, L and N represent capital, labour and human capital employed in the production of an aggregate output, while $\hat{K}$, $\hat{L}$ and $\hat{N}$ represent capital, labour and human capital employed in the production of knowledge (know-how) as signified by A, where $\dot{A} = dA/dt$ denotes the change of A per time unit. The three equations also need to be supplemented by a specification of the temporal motion of K, L and N as well as of $\hat{K}$, $\hat{L}$ and $\hat{N}$. When equation (13.1c) is excluded and (13.1b) is exogenously given, we have an aggregate regional production model of Solow type (Solow, 1994). Otherwise, the equation system represents a regional endogenous growth model (Romer, 1994). In either case, the model outlined in (13.1) tries to capture knowledge interaction and creation of all actors in a region into equations (13.1b) and (13.1c). In particular, it disregards knowledge flows to and from the region. A meaningful regional model has to consider interregional spillovers.

Sources of knowledge flows and knowledge creation
For firms in a functional region, we can identify the following three principal sources of knowledge:

- Knowledge creation based on R&D efforts in firms, research laboratories, universities and interactive regional innovation systems.
- Knowledge flows between actors inside a region, due to unintended local diffusion, knowledge interaction and inflow of knowledge embedded in new employees.
- Knowledge flows from sources outside the region based on imports, mediated interaction, foreign direct investment (FDI) and flows inside each (multinational) company group.

Studies that employ a knowledge production function relate a firm's knowledge inputs to its output of patent applications (for example Zucker et al., 1998). These studies indicate

that innovative firms in certain industries are highly dependent on knowledge generated by local university R&D (Feldman and Audretsch, 1999). The knowledge transfer and exchange may rely on a whole spectrum of mechanisms such as: (1) a flow of newly trained graduates from universities to industry; (2) technological spillovers from universities to industry; (3) industrial purchases of intellectual property of universities; (4) university researchers working as R&D consultants or serving on company boards; (5) university researchers leaving universities to work for industry or starting new firms; and (6) activities in incubator centres and science parks. In this context many studies emphasise the importance of localised knowledge flows (Varga, 1997; Andersson and Ejermo, 2005).

Recent studies compare a firm's accessibility to R&D activities in universities and in other firms, where accessibility here refers to the possibility to get in touch with R&D activities, and the associated contact costs. A clear outcome is that accessibility to other firms' R&D resources has the strongest impact on knowledge production (Andersson and Ejermo, 2005; Gråsjö, 2006). The results also imply that local knowledge flows are persistently important, whereas long-distance flows have very small impacts – except for firms belonging to a (multinational) company group, utilising the internal networks of its group to overcome long distances in the knowledge interaction. The pertinent subsidiaries can find locations in the proximity of places with specialised excellence (Dunning and Narula, 1995; Pfaffermayr and Bellak, 2002).

Why is knowledge spatially sticky?

In Arrow's (1962) contribution, it is observed that obstacles to communication may imply that technological as well as scientific knowledge may lose the property of being a pure public good. In fact, the most realistic assumption is that knowledge flows are affected by friction, reducing knowledge exchange.

In every particular case of knowledge transmission or transfer, the friction cost will vary because of geographic and other communication distances. Friction appears when knowledge is complex, as argued by Beckmann (1994, 2000), and when it is tacit, as described by Polanyi (1966). Therefore, knowledge will be 'sticky', in the words of von Hippel (1994). Face-to-face interaction is essential because it allows communicating agents to calibrate their coding, encoding and interpretation capabilities. Teece (1981) discusses new communication technologies that can modify the role of proximity in such calibrations. Knowledge exchangers may build knowledge networks by investing in intangible 'channels' or links for knowledge exchange (Beckmann, 1994). Such networks will reduce the importance of distance, but the investment costs will be recovered only to the extent that the knowledge links are used frequently. Evidently, we can partly explain the existence of multinational companies with their spatially distributed firm units by such network advantages (Kummerle, 1999; Narula, 2002). Antonelli et al. (2003) argue that stickiness of knowledge implies that knowledge can be shared by firms in a local environment of a functional region with little risk that the knowledge diffuses outside the region.

Let us finally observe that knowledge is an input to R&D activities and that the latter are carried out by firms, universities, laboratories and consultants that interact in innovation efforts. The above observations then imply that R&D activities will also be spatially sticky. In particular, it becomes relevant to describe and measure a firm's accessibility (potential of opportunities for interaction) to knowledge sources and to R&D activities of other actors. Studies by Andersson and Ejermo (2004, 2005) and, in

particular, Gråsjö (2006) show that the knowledge creation and innovation activities of a firm located in urban area (municipality) m, are primarily influenced by: (1) the local accessibility to knowledge sources inside the area, A_m^L; and (2) the regional accessibility to knowledge sources inside the functional region to which the area m belongs, A_m^R.

The two accessibility measures applied in Gråsjö (2006) have the following form:

$$A_m^L = G_m \exp\{-\lambda^L t_{mm}\} \tag{13.2a}$$

$$A_m^R = \sum_{s \in \hat{R}} G_s \exp\{-\lambda^R t_{ms}\} \tag{13.2b}$$

where G_s refers to knowledge resources in municipality s, t_{mm} is the average time distance between zones in municipality m, t_{ms} is the time distance between municipality m and s, $\hat{R}$ is the set of municipalities in the region to which m belongs – except municipality m itself. Moreover, λ^L is the distance discount factor inside a municipality and λ^R is intra-regional distance discount factor, where $\lambda^R > \lambda^L$. The finding in Gråsjö is that A_m^L and A_m^R influence the knowledge creation in m as two independent factors.

13.3 Knowledge in the regional macroeconomy

Can we claim that one region develops faster because of its superior knowledge assets? To what extent can a functional region be analysed as a separate entity, and which types of interregional influences must be taken into consideration?

The regional production function and endogenous growth

A natural starting point for gaining an increased theoretical understanding of the emerging knowledge economy is the new endogenous growth theory, which emphasises the role of the stock of accumulated knowledge and the growth of this stock. Endogenous growth models depict the growth process of an isolated economy and suggest that continuous increases in technological knowledge influence the aggregate economic growth, and they should perhaps be addressed as R&D models of economic growth (Romer, 1990).

The basic idea, as described earlier in (13.1), is that one part of an economy's resources (K, L, N) are used to produce an output that can be used for consumption as well as investment, while another part of the economy's resources ($\dot{K}$, $\dot{L}$, $\dot{N}$) are employed in the production of new technology. In such a single-region R&D model, the separation from the rest of the world becomes questionable. Regions trade with each other and technologies will diffuse from region to region, and this will impact upon each region's productivity variable, A. Moreover, new external knowledge will make parts of the 'old knowledge' obsolete in individual regions.

Lucas's growth model with endogenous human capital accumulation depicts how the gradual embodiment of knowledge in human beings (Lucas, 1988) is a driving force behind economic growth. Just like the contribution by Romer, Lucas recognises that economic growth is not emerging automatically as 'manna from heaven', but is the result of deliberate actions and choices of various stakeholders, including the government (Nijkamp, 2003).

The new versions of growth models shed light on Kaldor's (1963) famous statements: (1) per capita output continues to grow over time; (2) physical capital per worker grows over time; (3) the rate of return to capital is nearly constant; (4) the capital to output ratio remains approximately constant; (5) the proportion between labour and physical capital

income remains approximately unchanged over time; and (6) the growth rate differs substantially between countries (as between regions).

As observed in Griliches (1995) and in Barro and Sala-i-Martin (1995), the empirical research on the relation between R&D and economic growth is still embryonic. Griliches discusses three levels of identification: the firm, the industry and the national level. In all three cases cross-section studies yield the result that accumulated R&D is positively related to output per capita, whereas in time-series analyses the empirical findings are more ambiguous (Griliches, 1995).

Historically, growth accounting offers an approach to investigate empirically how growth rates depend on productivity-enhancing factors such as efficient resource allocation, scale factors (extent of market) and technology (Denison, 1962, 1967). Studies like Cheshire and Gordon (1998) and Cheshire and Magrini (2002) provide examples of regional growth accounting that are based on the concept of functional regions.

Interregional knowledge flows and multi-regional growth

In order to illuminate different aspects of interregional knowledge diffusion we shall make use of a model introduced in Andersson and Mantsinen (1980). The core of the model is a production function for each region r, $Q_r = Q(K_r, A_r)$, where K_r represents the production capital and A_r the accessible knowledge in the region. The knowledge resources in any region r is given by G_r, but the region can also benefit from knowledge G_s in other regions, denoted by s. The variable A_r summarises the compound effect of knowledge inside and outside the region.

The model components introduced above are sufficient to depict a landscape of interregional knowledge influences in a stylised model where the G variable may comprise scientific, technological and entrepreneurial knowledge as well as the capacity to develop new knowledge. Assume that this complex body of resources can be treated as a spatial public good, such that G_s in region s influences the accessible knowledge in region r via a distance-decay factor $f_{sr} = exp\{-\lambda t_{sr}\}$, where t_{sr} is the time distance between s and r. This yields

$$A_r = \sum_s f_{sr} G_s$$

reflecting total accessible knowledge in region r, where $f_{rr} > f_{sr} > 0$, for $s \neq r$. In this model, the accessible knowledge will change when any of the G variables is changed, and such changes are the outcome of R&D processes, depicted by the following differential equation:

$$\dot{G}_r = H(g_r \tau_r Q_r) \tag{13.3a}$$

where $dH/dz > 0$, where τ_r signifies the share of output, Q_r, that is allocated to knowledge-creation activities, and where g_r is the productivity of R&D in region r. A model of the multiregional development obtains, if we introduce a second dynamic equation that describes the accumulation of production capital, K_r, in the following way:

$$\dot{K}_r = s_r(1 - \tau_r) Q_r \tag{13.3b}$$

where s_r signifies the investment coefficient of region r. With suitable parameter values, there exists a long-run growth equilibrium, towards which the solution may approach over time. This process typically has a divergence phase followed by convergence. Moreover, improved accessibility spurs growth.

13.4 Knowledge creation in a region

Observing fundamental changes in contemporary societies, scholars have introduced concepts such as the information society, the service society, the post-industrial society and the knowledge society. We will refer to this as the C-society (Andersson, 1985a) when discussing the creative region and regional innovation systems.

Characteristics of the creative region

A classical statement by Schumpeter (1934) is that an innovation is the result of 'novelty by combination'. Such combinations have to be fuelled by creative individuals and groups of individuals. In Florida (2002, 2005) this aspect is emphasised by recognising the critical role of talented individuals, which implies that the concentration of innovations will be influenced by the location preferences of such individuals, belonging to 'the creative class'. In a Nordic context, the focus on creativity was introduced in the 1980s in a series of books by Andersson (1985a) and Andersson and Strömquist (1989). Two ideas are essential. First, it is suggested that the knowledge endowment of a region has the nature of a non-material infrastructure. Second, a model of individual creativity is introduced, emphasising seven abilities of the human brain. Beside the heuristic ability, these are: the ability to remember; to detect deep structures; to see and use ambiguity, multiplicity and variety; to appreciate paradoxes and surprises; to use disequilibria; and to use fundamental uncertainty.

Andersson (1985b, 1986) – as well as Castells (1989), Hall (1990) and Noyelle and Stanback (1985) – observes a clear path away from the industrial society, characterised by goods-handling activities, to a C-society characterised by knowledge-handling and development activities. In this transformation, the major driving force is creative activities, which generate new knowledge stimulated by culture and communication. Development, handling and presentation of new knowledge and information employs a steadily increasing share of the labour force, with strong spillover effects on industrial activities in manufacturing as well as in service production.

With the emergence of the C-society, industrial regions will lose their previous advantages and may have to restructure to regain prosperity. Instead, regions that afford creative milieux will benefit (Aydalot and Keeble, 1988). Moreover, urban regions with rich import networks and diversified import activities have a favourable position, because they acquire novelties from the world economy earlier than other regions.

Industries do not grow by expansion of existing activities, but through the emergence of new activities. Andersson (1985b) suggests that the associated creative processes are stimulated by: (1) tolerant attitudes towards experiment; (2) versatile composition of competences; (3) versatile basis for science, entrepreneurship and culture; (4) arenas for spontaneous and informal contact; (5) many-sided social and physical milieux; (6) perceptions that needs are greater than resources; and (7) a flexible social and economic organisation.

The fundamental role of metropolitan regions in the creative process depends upon their role as communication centres. Metropolitan regions are concentrations of

international communication in culture, business, politics and science. They also offer good opportunities to develop close-knit intra-regional communication networks within as well as between sectors of society. Metropolitan regions offer a physical proximity which facilitates the integration of multidisciplinary knowledge and improves the conditions for coping with uncertainty (cf. Patel and Pavitt, 1991).

Creative regions are also referred to as learning regions (Knight, 1995), emphasising a region's ability rapidly to generate, absorb and transform relevant knowledge and information as well as transform knowledge into learning. In addition, Keane and Allison (1999) suggest that learning regions are characterised by institutional thickness, based on relationships of trust and reciprocity and a mutual awareness of a common purpose. In Maillat and Kebir (2001) learning is a process to cope with permanent adaptation in the face of uncertainty.

Knowledge interaction in regional innovation systems

According to de la Mothe and Pacquet (1999), an innovation system consists of actors that interact in generation and exchange of economically useful information in a network with the following characteristics:

- Firms are part of a network of public and private sector organisations, whose activities and interactions initiate, import, modify and diffuse new technologies.
- The network includes both formal and informal linkages.
- There are flows of intellectual resources between organisations in the network.
- Learning is a key resource and key process.

When face-to-face interaction dominates, knowledge interaction will be framed by regional innovation systems, with informal routines, norms and institutions that are specific to each region (Andersson and Karlsson, 2006). Because of proximity externalities, a firm's innovation interaction becomes embedded in regional innovation systems (Johansson and Lööf, 2006). There is today a substantial agreement among scholars that the proximity gained by locating in large urban regions creates an advantage for innovative activities of firms by facilitating information and knowledge flows (Artle, 1959; Vernon, 1962; Andersson 1985a; Glaeser, 1999; Feldman and Audretsch, 1999).

Proximity externalities are mainly to be found within the borders of functional urban regions, as they define the geographical area within which frequent face-to-face interaction can take place (Johansson and Lööf, 2006). Various definitions of functional urban regions have been provided, for example, by Cheshire and Gordon (1995, 1998) for Europe, and by Johansson et al. (2002) for Sweden. In the latter case, a functional urban region is identified as a set of cities and towns between which the labour market commuting is mutually intense. Within functional regions, the average car travel-time between different urban areas is between 20 and 50 minutes, which matches the idea that knowledge spillovers are bounded in space.

How and to what extent do the regional innovation systems in various functional regions differ in their capacity to foster innovation activities by firms? First, they can differ in the amount, variety and richness of knowledge resources. Second, they can vary in terms of the intra-regional supply of knowledge-intensive labour, whose knowledge diffuses as they find new jobs over time. Third, functional regions vary in terms of their

Table 13.1 Classification of knowledge flows to a firm

Major types of knowledge flows	Examples of knowledge flows
Transaction-based flows	1. From knowledge providers selling knowledge inputs to a firm's R&D activities. 2. Innovations that are sold or licensed to a firm 3. Flows between firms that cooperate in R&D projects, where costs and benefits are regulated by explicit contracts.
Transaction-related flows	4. Knowledge embodied in the delivery of inputs from an input supplier to a firm. 5. Knowledge from an input supplier spills over unintentionally to an input-buying firm. 6. Knowledge from an input-buying firm spills over unintentionally to the input-selling firm.
Pure spillover flows	7. Unintentionally, knowledge spills over between firms in the same industry. 8. Unintentionally, knowledge spills over between firms belonging to different industries.

Source: Karlsson and Johansson (2006).

intra-regional accessibility to customers, suppliers and competitors as well as to knowledge providers such as universities, R&D laboratories, firms conducting R&D and consultancy firms (Henderson, 1974; Audretsch and Feldman, 1996; Karlsson and Johansson, 2006; Gråsjö, 2006).

The critical question is of course the multitude of options for interaction in regional innovation systems. Karlsson and Johansson (2006) present a classification of knowledge flows to a firm (Table 13.1). Knowledge flows can be transaction-based, transaction-related or pure knowledge spillovers. In the literature, there has been a focus on the pure knowledge spillovers. Initially much effort was put into studies of 'paper trails' (Jaffe et al., 1993), but over time the focus has shifted partly to the mobility of people (Zucker et al., 1998) and to knowledge management (Karlsson et al., 2004).

13.5 Knowledge, specialisation and growth

Knowledge patterns in large and small regions

Knowledge agglomeration is a self-reinforcing process, in which regions endowed with highly educated labour attract firms which benefit from good access to knowledge-intensive labour. As more firms move into these regions, the demand for amenities as well as for highly educated labour stimulates both investments in amenities and in-migration of highly educated labour. Thus, virtuous cycles might emerge, stimulating agglomeration of knowledge and creative potential (Glaeser and Kohlhase, 2004; Cheshire and Magrini, 2005). A well-established conclusion is that the proximity afforded by locations in large urban regions creates an advantage for firms by facilitating information and knowledge flows, following arguments presented in Artle (1959), Vernon (1962), Henderson (1974), Glaeser (1999) and Feldman and Audretsch (1999). This phenomenon may be classified as

proximity-based communication externality (Fujita and Thisse, 2002; Johansson and Quigley, 2004).

In most developed countries, a dominating share of all production takes place in the large urban regions. Most of the international and interregional trade takes place between these regions. Even more importantly, they are nodes in the international networks for knowledge and information transfer as well as milieux for creativity and innovation, and they comprise a diversified pattern of small-scale activities that relate to dynamic urbanisation economies in the sense of Jacobs (1984).

Small and medium-sized regions distinguish themselves by having limited local demand. Therefore, the local supply of distance-sensitive products (services) is not just smaller, but in particular less diversified. The second discriminating feature is that their knowledge resources are smaller. Furthermore, the share of labour with a long education is considerably smaller. Thus, these regions typically specialise in production based on natural resources, including tourist amenities. They can also have a base in local clusters that support production for export to markets in other regions. They are also candidates for relocation of decomposed subroutines as outlined in the following two subsections (Johansson and Karlsson, 2001).

Spatial product cycles

The product cycle theory is an attempt to develop a dynamic explanation to the division of labour between regions and nations (Vernon, 1966). In close association with the suggestions in the previous subsection, the product cycle theory assumes that firms introduce novel products with a higher than average frequency in urban regions with rich knowledge endowments and a creativity-stimulating milieu. The theory suggests, and empirical observations provide support to the idea, that those regions are large in most cases. There are three fundamental perspectives from which we can depict and analyse a product cycle. These are:

1. In the product cycle perspective, a product cycle model describes how product varieties in a novel product group increase their joint market share, usually at the expense of other established product groups, for which the market share reduces. The analysis of product cycle trajectories focuses on where the output of product groups originates and to which markets the sales are destined.
2. In the firm perspective, a product cycle refers to the temporal development of product varieties that a single firm supplies. The firm perspective is different from a region perspective, since a firm can have several different locations. Equally important, the single firm can over time initiate new product cycles, which implies that the firm rejuvenates itself by phasing out 'obsolete' product varieties, while simultaneously introducing new varieties which form new product cycles. Such a firm may develop its new products in one type of region and produce products with a declining market trend in another type of region.
3. In the region perspective, certain regions may host the supply of young product varieties, while other regions persistently offer locations for the supply of product varieties with stagnating or declining market shares.

Combining the first and the third perspective, a product cycle model can be viewed as a dynamic complement to the classical theory of comparative advantage. It involves three

types of industrial dynamics: (1) technological development, which introduces new products; (2) introduction of new or improved production processes; and (3) changes of the market organisation with new market channels.

The product cycle model provides a stylised framework for understanding the changes in the demand for different types of inputs over the life cycle of a product. The trajectory of a life cycle can be described with regard to an entire product group, comprising varieties that satisfy similar needs. A life cycle trajectory can also refer to the development of a firm's supply of one or several such varieties. When the focus is on a product group, we may recognise the juvenile stage of such a group when all its varieties are young, non-standardised objects, which are still in a process of experimental design. In this early stage of development, the R&D work benefits from taking place in a creative milieu, with a rich supply of knowledge resources accessible to the design activities (Vernon, 1966; Norton and Rees, 1979; Malecki, 1981; Nijkamp, 1986).

Given that the design and market penetration process is successful, the product group enters a phase when the output and sales of the new product varieties expand at a fast rate. In this phase, the pertinent firms have better opportunities to routinise the production, distribution and marketing activities, and this can further stimulate the expansion. The routinisation of firm operations is facilitated when the design of product varieties is standardised. Then, the unit cost of each variety can be reduced, which will stimulate market expansion. In this stage, the location may shift to places that are less knowledge-intensive, generating outsourcing and offshoring.

Knowledge intensity and location dynamics

Product standardisation and process routinisation are key notions in the model of product cycle dynamics developed by Johansson and Andersson (1998). Along a product cycle path the knowledge intensity is high when a product is non-standardised and the production process is non-routinised. Standardisation and routinisation imply reduced knowledge intensity, favouring alternative locations (Johansson and Karlsson, 1986).

In order to make the location dynamics transparent, we shall compare costs in regions s and k, where s is more and k less knowledge rich. The market demand in the early phase is assumed to stimulate the supply to reach the volume x. The production requires basic resources (b-type) and knowledge resources (g-type). The total amount of b-resources is given by $B(x) = F_b + bx$, where F_b denotes fixed and bx variable resources. $G(x) = F_g + gx$ denotes the total amount of knowledge resources, where F_g denotes fixed and gx variable resources.

Region-specific prices of basic resources in region s and k are denoted by ρ_s and ρ_k, while region-specific prices of knowledge resources are denoted by ω_s and ω_k. Region s is knowledge rich, and we assume that $\rho_s > \rho_k$ and $\omega_s \leq \omega_k$, which implies that basic resources have a lower price in region k.

In the early stage of the product cycle, region s has a location advantage when the following inequality applies:

$$C_s(x) = \rho_s B(x) + \omega_s G(x) < C_k(x) = \rho_k B(x) + \omega_k G(x) \tag{13.4a}$$

As the product cycle develops the output increases to x^*, the fixed resource requirements change to F_g^* and F_b^*, and the variable input coefficients change to g^* and b^*. At this mature

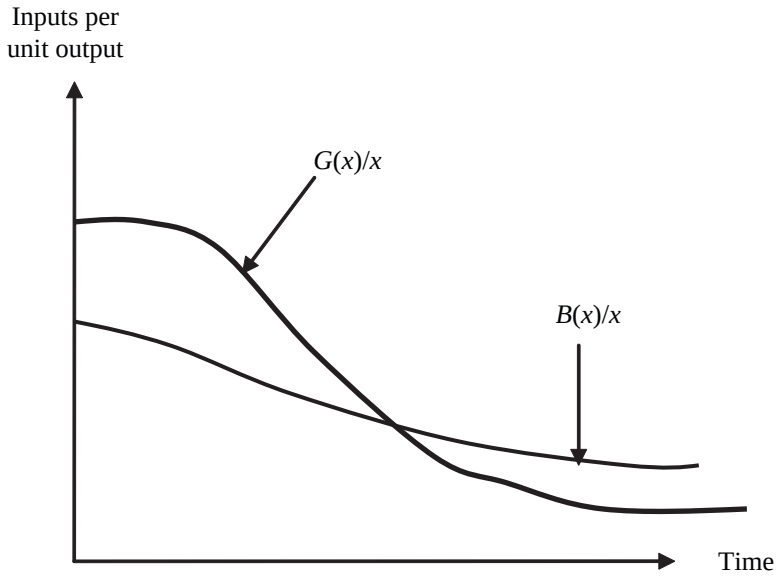


Figure 13.2 Knowledge resources per unit output fall faster than basic resources per unit output

stage we have $G^*(x^*) = F^*_g + b^*x^*$ and $B^*(x^*) = F^*_b + b^*x$. According to product cycle assumptions, the technology satisfies the following two conditions: $F_g/F_b > F^*_g/F^*_b$, and $g/b > g^*/b^*$. Therefore, the location advantages may shift in the mature phase of technology such that region s becomes a less favourable location:

$$C^*_s(x^*) = \rho_s B^*(x^*) + \omega_s G^*(x^*) > C^*_k(x) = \rho_k B^*(x^*) + \omega_k G^*(x^*) \qquad (13.4b)$$

The condition in (13.4b) means that in the mature phase of development, knowledge resources play a less important role than in earlier phases. As the product cycle develops, the technology will change and this can imply that both $B(x_t)/x_t$ and $G(x_t)/x_t$ fall as time, t, increases. However, if $G(x_t)/x_t$ falls more than $B(x_t)/x_t$, then a shift in location advantage may occur as shown in (13.4b) and illustrated in Figure 13.2.

Students of product cycles have often argued that in mature stages firms start to employ large-scale routines, implying that F_b increases with maturity (Utterback and Abernathy, 1975; Andersson and Johansson, 1984; Klepper, 1996). However, a more relevant hypothesis is that the individual firm decomposes its routines in mature phases, giving the firm opportunities to outsource and offshore several of the decomposed routines into a supply chain network. First, each such outsourced routine activity can be located in a region with low costs of basic resources. Second, the size of such units may not be extremely large. At the same time, the entire supply chain can still represent a large-scale operation.

Consider now an outsourced production unit executing a subroutine operation in a low-cost location. It seems likely that such a production unit will give priority to process innovations in line with model predictions in Klepper (1996) or Andersson and Johansson (1984). At the same time, the 'headquarter firm' may focus on development of new

product varieties to create product innovations. In such a scenario, a multi-unit firm would have firm units in both knowledge-intensive regions and in regions with a comparative advantage in routine-like activities.

Innovation studies that examine the role of corporate structure seem to provide support to the conclusion that multinational firms remain R&D-intensive in their respective home country, where they continue to generate product innovations, while at the same time locating production units in less knowledge-intensive regions around the world (Freeman, 1992; Criscuolo et al., 2005; Ebersberger and Lööf, 2005). Such innovation-persistent enterprises continue to keep a large share of their innovation activities in knowledge-intensive regions in their home country. Such behaviour also helps to stimulate these regions to remain rich in knowledge resources and to maintain their creative milieux.

13.6 Conclusions for regional development policies

Accessibility to knowledge-intensive labour obtains for firms in regions which are capable of attracting households that supply this labour. Hence, the regional consumption and cultural milieu as well as regional amenities of other kinds are crucial features. Such demand-based advantages relate to urbanisation economies, providing households with diversified consumption opportunities and firms with diversified demand.

Location advantages evolve slowly in path-dependent processes. This is especially true for knowledge-based advantages. To be successful, regional policy therefore has to focus on structural adjustments of tangible and non-tangible infrastructure. Universities and university colleges are agents of human capital formation and may support enhancement of local knowledge assets, while various non-profit organisations and similar institutions may catalyse the formation of social capital.

In view of the arguments put forward in this chapter, it is possible to identify four areas for regional policies that relate to a region's knowledge resources:

- Knowledge policies, focusing on education and training of the labour force, development of innovation systems that support R&D, patenting, product and commercial innovations, and improving the capacity to absorb externally diffused knowledge.
- Household milieu policies, influencing life conditions by forming human and social capital, and enriching households' opportunities with regard to recreation, job accessibility and natural environment attributes. Knowledge workers are far more demanding in these regards than the labour force on average.
- Facility policies, comprising built infrastructure for urban life, transport, Internet and telecommunications, property development, urban management including transport demand, and land value mechanisms.
- Firm milieu policies, stimulating technology diffusion, facilitating supply of venture capital, supporting firm start-ups, attracting direct investments by external firms, orchestrating cluster formation and improving conditions for labour market adjustments.

References

Andersson, Å.E. (1985a), *Kreativitet – Storstadens framtid*, Stockholm: Prisma.
Andersson, Å.E. (1985b), 'Creativity and regional development', *Papers of the Regional Science Association*, **56**, 5–20.

Andersson, Å.E. (1986), 'The four logistical revolutions', *Papers of the Regional Science Association*, **59**, 1–12.
Andersson, Å.E. and B. Johansson (1984), 'Knowledge intensity and product cycles in metropolitan regions', WP-84-13, IIASA, Laxenburg.
Andersson, Å.E. and J. Mantsinen (1980), 'Mobility of resources: accessibility of knowledge and economic growth', *Behavioural Science*, **25**, 353–66.
Andersson, Å.E. and U. Strömqvist (1989), *K-samhällets framtid*, Stockholm: Prisma.
Andersson, M. and O. Ejermo (2004), 'Sectoral knowledge production in Swedish regions 1993–1999', in C. Karlsson, P. Flensburg and S.-Å. Hörte (eds), *Knowledge Spillovers and Knowledge Management*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, 143–70.
Andersson, M. and O. Ejermo (2005), 'How does accessibility to knowledge sources affect the innovativeness of corporations? Evidence from Sweden', *Annals of Regional Science*, **39**, 741–65.
Andersson, M. and C. Karlsson (2006), 'Regional innovation systems in small and medium-sized regions', in B. Johansson, C. Karlsson and R. Stough (eds), *The Emerging Digital Economy. Entrepreneurship, Clusters and Policy*, Berlin: Springer, pp. 55–81.
Antonelli, C., R. Marchionatti and S. Usai (2003), 'Productivity and external knowledge: the Italian case', *Rivista Internazionale di Scienze Economiche e Commerciali*, **50**, 69–90.
Arrow, K.J. (1962), 'Economic welfare and the allocation of resources for invention', in R. Nelson (ed.), *The Rate and Direction of Inventive Activity*, NBER, Princeton: Princeton University Press, pp. 619–22.
Artle, R. (1959), *The Structure of the Stockholm Economy – Toward a Framework for Projecting Metropolitan Community Development*, Business Research Institute, Stockholm School of Economics, Stockholm.
Audretsch, D.B. and M.P. Feldman (1996), 'R&D spillovers and the geography of innovation and production', *American Economic Review*, **86**, 630–40.
Aydalot, P. and D. Keeble (1988), *High-technology Industry and Innovation Environments: The European Experience*, London: Routledge and Kegan Paul.
Barro, R.J. and X. Sala-i-Martin (1995), *Economic Growth*, New York: McGraw-Hill.
Beckmann, M.J. (1994), 'On knowledge networks in science: collaboration among equals', *Annals of Regional Science*, **28**, 233–42.
Beckmann, M.J. (2000), 'Interurban knowledge networks', in D. Batten (ed.), *Learning, Innovation and Urban Evolution*, London: Kluwer Academic, pp. 127–35.
Castells, M. (1989), *The Informational City*, Oxford: Blackwell.
Chambers, R.G. (1988), *Applied Production Analysis: A Dual Approach*, Cambridge: Cambridge University Press.
Cheshire, P. and I. Gordon (1995), *Territorial Competition in an Integrating Europe*, Aldershot: Avebury.
Cheshire, P. and I. Gordon (1998), 'Territorial competition: some lessons for policy', *Annals of Regional Science*, **32**, 321–46.
Cheshire, P. and S. Magrini (2002), 'Counteracting the counterfactual: new evidence on the impact of local policy from the residuals', in B. Johansson, C. Karlsson and R.R. Stough (eds), *Regional Policies and Comparative Advantage*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, pp. 209–38.
Cheshire, P. and S. Magrini (2005), 'Regional demographic or economic dynamism? Different causes, different consequences', Paper presented at the international workshop Innovation, Dynamic Regions and Regional Dynamics, 16–18 June, Jönköping, Sweden.
Cornes, R. and T. Sandler (1986), *The Theory of Externalities, Public Goods and Club Goods*, Cambridge: Cambridge University Press.
Criscuolo, P., R. Narula and B. Verspagen (2005), 'The role of home and host country innovation systems in R&D internationalisation: a patent citation analysis', *Economics of Innovation and New Technology*, **14**, 417–33.
de la Mothe, J. and G. Pacquet (1999), *Local and Regional System of Innovation*, Boston: Kluwer, Dordrecht OBSERVE.
Denison, E.F. (1962), 'The sources of economic growth in the United States and the alternatives before us', Committee for Economic Development, New York.
Denison, E.F. (1967), *Why Growth Rates Differ: Post-War Experience in Nine Western Countries*, Washington, DC: Brookings Institution.
Dosi, G., C. Freeman, R. Nelson, G. Silverberg and L. Soete (eds) (1988), *Technical Change and Economic Theory*, London: Pinter Publishers.
Dunning, H.H. and R. Narula (1995), 'The R&D activties of foreign firms in the United States', *International Studies of Management and Organization*, **25**, 39–73.
Ebersberger, B. and H. Lööf (2005), 'Innovation behaviour and productivity performance in the Nordic region: does foreign ownership matter?', CESIS Working Paper No. 27, Centre for Science and Innovation Studies, Royal Institute of Technology, Stockholm.
Feldman M.P. and D.B. Audretsch (1999), 'Innovation in cities: scienced-based diversity, specialization and localized competition', *European Economic Review*, **43**, 409–29.
Florida, R. (2002), *The Rise of the Creative Class*, New York: Basic Books.

Florida, R. (2005), *Cities and the Creative Class*, New York: Routledge.
Freeman, C. (1992), 'Formal scientific and technical institutions in the national system of innovation, in B.-Å. Lundvall (ed.), *National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning*, London: Pinter Publishers, pp. 169–87.
Fujita, M. and J.F. Thisse (2002), *Economics of Agglomeration: Cities, Industrial Location and Regional Growth*, Cambridge: Cambridge University Press.
Glaeser, E. (1999), 'Learning in cities, *Journal of Urban Economics*, **100**, 254–77.
Glaeser, E.L. and J.E. Kohlhase (2004), 'Cities, regions and the decline of transport costs', *Papers in Regional Science*, **83**, 197–228.
Griliches, Z. (1995), 'R&D and productivity: economic results and measurement issues', in P. Stoneman (ed.), *Handbook of the Economics of Innovation and Technological Change*, Oxford: Blackwell, pp. 52–89.
Grossman, G.M. and E. Helpman (1991), 'Quality ladders and product cycles', *Quarterly Journal of Economics*, **106**, 557–86.
Gråsjö, U. (2006), 'Spatial spillovers of knowledge production: an accessibility approach', JIBS Dissertation Series No. 034, Jönköping International Business School, Sweden.
Hall, P. (1990), 'High-technology industry and the European scene', in *Urban Challenges* (SOU 1990:33), Statens offentliga utredningar, Stockholm, pp. 117–33.
Henderson, J.V. (1974), 'The size and type of cities, *American Economic Review*, **89**, 640–56.
Jacobs, J. (1984), *Cities and the Wealth of Nations*, New York: Random House.
Jaffe, A.B., M. Trajtenberg and R. Henderson (1993), 'Geographical localisation of knowledge spillovers as evidenced by patent citations', *Quarterly Journal of Economics*, **108**, 577–98.
Johansson, B. and Å.E. Andersson (1998), 'A Schloss Laxenburg model of product cycle dynamics', in M.J. Beckmann, B. Johansson, F. Snickars and R. Thord (eds), *Knowledge and Networks in a Dynamic Economy*, Berlin: Springer, pp. 181–219.
Johansson, B. and C. Karlsson (1986), 'Industrial application of information technology: speed of introduction and labour force competence', in P. Nijkamp (ed.), *Technological Change, Employment and Spatial Dynamics*, Berlin: Springer, pp. 401–28.
Johansson, B. and C. Karlsson (2001), 'Geographic transaction costs and specialisation opportunities of small and medium-sized regions: scale economies and market extension', in B. Johansson, C. Karlsson and R.R. Stough (eds), *Theories of Endogenous Regional Growth: Lessons for Regional Policies*, Berlin: Springer, pp. 150–80.
Johansson, B. and J. Quigley (2004), 'Agglomeration and networks in spatial economies', *Papers in Regional Science*, **83**, 165–76.
Johansson, B. and H. Lööf (2006), 'Innovation activities explained by firm attributes and location', CESIS Working Paper No. 63, Centre of Excellence for Science and Innovation Studies, Royal Institute of Technology, Stockholm.
Johansson, B., J. Klaesson and M. Olsson (2002), 'Time distances and labour market integration', *Papers in Regional Science*, **81**, 305–27.
Kaldor, N. (1963), 'Capital accumulation and economic growth', in F.A. Lutz and D.C. Hague (eds), *The Theory of Capital: Proceedings of a Conference Held by the International Economic Association*, London: Macmillan, pp. 177–222.
Karlsson, C. and M. Andersson (2006), 'The location of industry R&D and the location of university R&D: how are they related?', in C. Karlsson et al. (eds), *Innovation, Dynamic Regions and Regional Dynamics*, Berlin: Springer.
Karlsson, C. and B. Johansson (2006), 'Dynamics and entrepreneurship in a knowledge-based economy', in C. Karlsson, B. Johansson and R.R. Stough (eds), *Entrepreneurship and Dynamics in the Knowledge Economy*, New York: Routledge, pp. 12–46.
Karlsson, C., P. Flensburg and S.-Å. Hörte (2004), 'Introduction: knowledge spillovers and knowledge management', in C. Karlsson, P. Flensburg and S.-Å. Hörte (eds), *Knowledge Spillovers and Knowledge Management*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, pp. 3–31.
Keane, J. and J. Allison (1999), 'The interaction of the learning region and local and regional economic development: analysing the role of higher education', *Regional Studies*, **33**, 896–902.
Klepper, S. (1996), 'Exit, entry, growth and innovation over the product life cycle', *American Economic Review*, **86**, 562–83.
Knight, R.V. (1995), 'Knowledge based development: policy and planning implications for cities', *Urban Studies*, **32**, 225–60.
Kobayashi, K. (1995), 'Knowledge networks and market structure: an analytical perspective', in D.F. Batten, J. Casti and R. Thord (eds), *Networks in Action. Communication, Economics and Human Knowledge*, Berlin: Springer, pp. 127–58.
Kobayashi, K. and Å.E. Andersson (1994), 'A dynamic input–output model with endogenous technical change', in B. Johansson, C. Karlsson and L. Westin (eds), *Patterns of a Network Economy*, Berlin: Springer, pp. 243–59.

Kummerle, W. (1999), 'Foreign direct investment in industrial research in the pharmaceutical and electronics industries: results from a survey of multinational firms', *Research Policy*, **28**, 179–93.
Lucas, R.E. (1988), 'On the mechanics of economic development', *Journal of Monetary Economics*, **22**, 3–42.
Maillat, D. and L. Kebir (2001), 'The learning region and territorial production systems, in B. Johansson, C. Karlsson and R.R. Stough (eds), *Theories of Endogenous Regional Growth. Lessons for Regional Policies*, Berlin: Springer, pp. 255–77.
Malecki, E.J. (1981), 'Product cycles and regional economic change', *Technological Forecasting and Social Change*, **19**, 291–306.
Narula, R. (2002), 'Innovation systems and "inertia" in R&D location: Norwegian firms and the role of systemic lock-in', *Research Policy*, **31**, 795–816.
Nelson, R.R. and S.G. Winter (1982), *An Evolutionary Theory of Economic Change*, Cambridge, MA: Harvard University Press.
Nijkamp, P. (ed.) (1986), 'Technological change, employment and spatial dynamics', *Lecture Notes in Economics and Mathematical Systems*, Vol. 270, Berlin: Springer-Verlag.
Nijkamp, P. (2003), 'Entrepreneurship in a modern knowledge economy', *Regional Studies*, **37**, 395–405.
Norton, R. and J. Rees (1979), 'The product life cycle and the decentralisation of American manufacturing', *Regional Studies*, **13**, 48–90.
Noyelle, T.J. and T.M. Stanbeck (1984), *The Economic Transformation of American Cities*, Totowa: Rowman & Allanheld.
Patel, P. and K. Pavitt (1991), 'Large firms in the production of the world's technology: an important case of "non-globalisation" ', *Journal of International Business Studies*, **22**, 1–21.
Pfaffermayr, M. and C. Bellak (2002), 'Why foreign-owned firms are different: a conceptual framework and empirical evidence from Austria', in R. Jungnickel (ed.), *Foreign-Owned Firms: Are They Different?*, London: Palgrave Macmillan, pp. 13–57.
Polanyi, M. (1966), *The Tacit Dimension*, London: Routledge & Kegan Paul.
Romer, P. (1990), 'Endogenous technological change', *Journal of Political Economy*, **98**, 71–102.
Romer, P. (1994), 'The origins of endogenous growth', *Journal of Economic Perspectives*, **8**, 3–32.
Schumpeter, J.A. (1934), *The Theory of Economic Development*, Cambridge, MA: MIT Press.
Solow, R.M. (1994), 'Perspectives of growth theory', *Journal of Economic Perspectives*, **8**, 45–54.
Teece, D.J. (1981), 'The market for know-how and the efficient international transfer of technology', *Annals of the American Association of Political and Social Sciences*, **458**, 54–67.
Utterback, J.M. and W.J. Abernathy (1975), 'A dynamic model of process and product innovation', *OMEGA*, **3**, 639–56.
Varga, A. (1997), *University Research and Regional Innovation: A Spatial Econometric Analysis of Academic Technology Transfers*, Boston, MA: Kluwer Academic Publishers.
Vernon, R. (1962), *Metropolis 1985*, Cambridge, MA: Harvard University Press.
Vernon, R. (1966), 'International investment and international trade in the product cycle', *Quarterly Journal of Economics*, **80**, 190–207.
von Hippel, E. (1994), 'Sticky information and the locus of problem solving: implications for innovation', *Management Science*, **40**, 429–39.
Zucker, L.G., M.R. Darby and J. Armstrong (1998), 'Geographically localised knowledge: spillovers or markets?', *Economic Inquiry*, **36**, 65–86.

14 Agglomeration externalities, innovation and regional growth: theoretical perspectives and meta-analysis

Henri L.F. de Groot, Jacques Poot and Martijn J. Smit

14.1 Introduction

Since the mid-1980s there has been a remarkable revival of research devoted to both theoretical modelling and empirical verification of the causes of long-run economic growth at spatial scales ranging from the global economy down to the local community (see, for example, Barro and Sala-i-Martin, 2004, for an overview of the field). One of the major drivers of this research activity was the realisation that growth cannot be understood without investigating the characteristics, geography, causes and consequences of innovation – namely, the implementation of new or significantly improved products, processes, business practices, workplace organisation or external relations (OECD, 2005). Innovation takes place in dynamically diverse, geographically concentrated and imperfectly competitive spaces that can only be analysed by abandoning conventional assumptions of perfectly competitive markets and constant returns to scale. This realisation led to the development of 'new' theories of growth, economic geography, trade and industrial organisation (see, for example, Krugman, 1995; Brakman and Heijdra, 2004).

In the knowledge-driven globally connected regional economy, agglomeration forces that rely on proximity continue to increase in importance. This occurs paradoxically despite declining real costs of information, communication and transportation. The relevance of agglomeration is revealed by the continuing urbanisation of the global population. About half the world population now lives in cities and this is expected to increase further to 60 per cent by 2030 (UNFPA, 2007). Although the number of 'world cities' with populations of more than 10 million inhabitants continues to increase, global urbanisation is primarily due to the growth of smaller cities of up to 500 000 inhabitants. While mega cities have hugely diverse economies, smaller cities may find a niche in specialised economies or clusters of connected activities (see, for example, Fujita and Thisse, 2002).

Understanding the existence and growth of mega and smaller cities and their surrounding hinterlands – that together make up functional regions – requires consideration of a wide range of factors that have been elaborated in the above-mentioned 'new' theories of innovation and growth and that have been empirically tested in a large range of studies around the world.[1] The growth of cities results from a complex chain that starts with scale: endowments of labour, capital and knowledge. The productivity of the open urban economy also depends on spatial factors, internally through density and infrastructure and externally through spatial interaction with other cities and regions. Resources, production factors and geography then combine with an industrial structure characterised by specialisation, competition and diversity to yield innovation and productivity growth that encourages employment expansion.

In the long run, new jobs can only be filled through natural increase of the urban population or through net inward migration. Given that rising real incomes in cities lead to lower fertility, urban population growth is in practice primarily driven by inward migration of workers who are often positively self-selected in terms of entrepreneurial abilities and skills. In the presence of economic diversity and increasing returns, capital and labour do not flow in opposite directions, as in static neoclassical theory. Instead, the city attracts capital too. Many aspects of this self-reinforcing and virtuous process yield benefits that are external to individual market transactions and such externalities are therefore central to agglomeration processes (see Fujita and Thisse, 2002).

This chapter revisits the issue of the importance of externalities that have provided alternative explanations for innovation and urban growth. Following the seminal contribution by Glaeser et al. (1992), a large volume of empirical research has tried to identify the roles of industrial concentration and specialisation (Marshall–Arrow–Romer – MAR – externalities, originally noted by Marshall, 1890), economic and social diversity leading to cross-sectoral spillovers (Jacobs externalities, after Jacobs, 1969), and the intensity of competition (Porter externalities, after Porter, 1990). However, this research endeavour has only been partially successful. Glaeser (2000) concluded that the relative importance of such externalities remains largely unresolved. In their review of growth, development and innovation, Cheshire and Malecki (2004, p. 263) additionally noted that 'an important element in any research agenda is a job of synthesis'.

In this chapter we therefore evaluate the statistical robustness of evidence for agglomeration externalities by means of a form of quantitative literature review, commonly referred to as meta-analysis, of 31 published articles that provide empirical evidence on the impact of agglomeration and innovation on the growth of cities. Meta-analysis is becoming increasingly popular in economics after having a longer tradition in biomedical and behavioural sciences.[2] The analysed articles yield 393 indicators of the statistical significance of agglomeration externalities on growth. These so-called effect sizes are linked to study characteristics by means of an ordered probit analysis. The evidence in the literature on the role of the specific externalities is rather mixed, although for each type of externality we can identify clearly how various aspects of primary study design, such as the adopted proxy for growth, the data used, and the choice of covariates, influence the outcomes. We find most clear-cut evidence for a positive effect of diversity, supporting Jacobs's view. Somewhat less conclusive evidence was found for a positive impact of competition on city growth. Regarding the effect of specialisation, the evidence is largely mixed.

In the next section we review theoretical perspectives on the nature of agglomeration externalities and their impact on growth and development. From this literature, several testable hypotheses can be derived. We subsequently turn in section 14.3 to the empirical literature that has investigated the impact of agglomeration externalities. Central to this review is the approach adopted in the seminal paper by Glaeser et al. (1992), which has triggered the research agenda in this area and therefore deserves a relatively more detailed review than other contributions. In section 14.4, we provide a statistically based description of the available evidence using tools developed under the heading of meta-analysis. The final section sums up and suggests ways in which this literature can be fruitfully developed further from here on.

14.2 Theoretical perspectives on agglomeration externalities and growth

Considerable effort has been devoted in recent years to modelling the nature and impact of agglomeration (for example Fujita and Thisse, 2002). While some of these ideas go back to Marshall (1890), Christaller (1966 [1933]), Ohlin (1933) and Lösch (1954 [1940]), agglomeration continues to attract attention because of the continuing urbanisation throughout the world noted earlier and the complexities of defining and measuring agglomeration effects.

Historically, agglomerations of economic activity resulted from the efficiency and strategic advantage of settlement at specific locations, usually determined by geography (access to water, other resources and the features of the landscape) and the interrelated development of trade routes. The benefits of such spatial concentration of economic activity in which all economic agents benefit from lower transaction and coordination costs are referred to as localisation externalities.

Other types of externalities are those of urban scale and density. An increase in population increases aggregate demand and enables firms to expand output without efficiency or productivity improvements. In this respect, scale and density are interrelated but not identical. A greater scale of activity may be accommodated by increasing urban sprawl at constant density, while alternatively the current tendency for a return of knowledge workers to the inner city may increase urban core density without changing scale. Scale and density effects may be referred to as urbanisation externalities. The importance of these may be gauged from the ease with which, through demand effects, cities can absorb large numbers of immigrants over a very short period of time (such as in the Mariel boatlift, the rapid exodus of an estimated 125 000 Cubans to Florida in the summer of 1980; see, for example Bodvarsson et al., 2007). A fiscal externality also exists in that public goods can be funded through a lower per capita lump-sum tax when the urban population increases. On the other hand, urbanisation externalities can also be negative and determine the limits to urban growth through pollution and congestion effects with respect to infrastructure and land use (for example Glaeser, 1998).

Glaeser et al. (1992) refer to the above externalities as static in that they explain the cross-sectional distribution of economic activity, levels of productivity and amenities, but not changes in sector-specific productivity due to, for example, knowledge spillovers. The latter are referred to as knowledge externalities and these dynamic externalities are the focus of the empirical analysis of Glaeser et al. (1992) as well as the analysis in the present chapter.

To provide a basic framework for analysis, we will now turn to an illustration of the main dependencies between inputs, productivity and utility using a simple model. We will then proceed to relate the analysis by Glaeser et al. (1992) to this framework. Most modern modelling of economic development starts from a general equilibrium perspective in which profit-maximising firms in any given region and sector determine output and inputs based on the productivity of resources and given factor prices.[3] Specifically:

$$w_{irt} = \pi_{it} MPL(L_{firt}, K_{firt}, A_{firt}) \tag{14.1}$$

$$\rho_t = \pi_{it} MPK(L_{firt}, K_{firt}, A_{firt}) \tag{14.2}$$

in which f indexes the firm $(1,2,..,F_{irt})$, i indexes the industry $(1,2,\ldots,I)$, r indexes the region $(1,2,..,R)$, t is a time index. The variable w_{irt} refers to the wage paid to workers in industry

i, region *r* at time *t* (each firm in a given local labour market pays the 'going' wage),[4] π_{it} refers to the price of a product (assumed to be traded in global markets so that it is equal for each firm and region), ρ_t is the price of capital (which is equal everywhere due to the assumption of perfect international and intersectoral capital mobility), L_{firt} refers to the labour input, K_{firt} refers to the capital input, A_{firt} refers to the knowledge input, and *MPL* and *MPK* refer to the physical marginal products of labour and capital, respectively, which are functions of the inputs. These functions have the usual partial derivatives, that is, $MPL_L < 0$, $MPL_K > 0$, $MPL_A > 0$, $MPK_L > 0$, $MPK_K < 0$ and $MPK_A > 0$. Capital is perfectly mobile and allocated such that the rate of return is equalised across sectors and regions. Workers are also perfectly mobile such that utility is equalised across space, and wage differentials reflect amenity differentials. Hence, the utility of a worker in industry *i* and region *r* can be described as:

$$U_{irt} = \overline{U}_{it} = \phi(w_{irt}, Q_{rt}) \tag{14.3}$$

in which workers of industry *i* reach the same utility $\overline{U}_{it}$ in every region, with Q_{rt} referring to the amenity level in region *r*. Combining this supply side with demand equations for the *I* commodities, and with given nationwide factor endowments, an equilibrium can in principle be determined.

In order to study the dynamics of such an economy, it is clear that the neoclassical model developed by Solow (1956) and Swan (1956) of long-run growth in which the long-run steady state is determined by a given technology, by investment funded from local savings and by natural increases in the workforce, is not appropriate. Among the most important problems is the fact that we have an open system in which capital accumulation and spatial reallocation of workers depend on the development of knowledge across all regions. The long-run tendency of such a system depends on the endogeneity of technological change and the nature of the spatial interaction and spillovers (for example, Nijkamp and Poot, 1998).

First, we can consider the growth in knowledge at the level of the firm. As in agent-based modelling (for example Zhang, 2003), the micro-level employment response of employers, following a productivity increase, determines one side of the motion in the system. Formally, the productivity growth can be described by:

$$A_{firt+1} = A_{firt} + \Delta A_{firt}(t, \mathbf{L}_t, \bar{A}_t) \tag{14.4}$$

in which ΔA_{firt} refers to the shift in the firm's knowledge input, which is a function of time (t), the distribution of employment across firms, industries and regions at time *t* represented by the three-dimensional array $\mathbf{L}_t$ with elements L_{firt} representing employment by individual firms in that industry-region at time *t*, and the economy-wide level of knowledge $\overline{A}_t$. The arguments of the function ΔA_{firt} are external to firm *k*'s actions, except for L_{kirt}. The partial derivatives of ΔA_{firt} with respect to *t* and $\overline{A}_t$ are positive. The first of these relates to exogenous long-run technological change and the second to the economy-wide benefits of, for example, a high level of education.[5] There are theoretically several mechanisms through which the array $\mathbf{L}_t$, the configuration of employment across firms, industries and regions, can affect productivity growth. These include the MAR, Porter and Jacobs externalities referred to earlier. The extent to which any of these externalities, or a

combination, has a statistically significant impact on productivity growth (or its proxy) is the primary objective of the meta-analysis to which we turn later in this chapter.

However, the actual employment decision of the firm is also a function of another set of externalities, namely those that affect the utility of workers (for example Glaeser, 1998). These can be positive or negative. Positive externalities of urban growth include the benefits of urban amenities, the enjoyment of cultural diversity and the fiscal externality of larger local tax revenues that enable lower local tax rates or higher quality recreational amenities (for example Florida, 2002). Negative externalities of urban growth include congestion, pollution, a decline in social cohesion and an increase in social problems. Formally:

$$Q_{rt+1} = Q_{rt} + \Delta Q_{rt}(t, \mathbf{L}_t, \Delta K_{rt}, \overline{A}_{rt}) \tag{14.5}$$

The partial derivatives of the function ΔQ_{rt} may be expected to be negative with respect to t (depreciation of existing amenities), positive or negative with respect to regional investment ΔK_{rt} (dependent on whether this generates more amenities and infrastructure, or disutility, for example through visual pollution), and positive with respect to the local overall level of knowledge $\overline{A}_{rt}$ (education may reduce crime and improve social cohesion). It is hard to say a priori how a change in the array $\mathbf{L}_t$ would affect the quality of life in region r. Greater employment in 'clean' service sector firms might improve the quality of life, whereas greater employment overall may generate pollution and congestion externalities. On balance, we are assuming a negative net amenity externality of city output growth, which is consistent with much of the available empirical evidence (see, for example, Capello, 2004). This implies that nominal wages must increase to compensate. The net effect on employment depends on the compensating growth in nominal wages. If the negative externality effect is relatively minor, the firm's employment will increase. If the negative externalities are significant, firms can only attract workers when the offered wage increases substantially and employment will decline.

In order to describe the dynamics of the multiregional system in the simplest possible way, we consider a two-region case in which one of the two regions is affected by such positive and negative agglomeration externalities. The adjustments along the equilibrium growth path are illustrated in Figure 14.1.

The top half of the figure depicts the impact of the positive production externality. The bottom half depicts the impact of a negative utility externality. The left side depicts the agglomerated region (region 1) and the right side a region without agglomeration-linked externalities (region 2). The demand curves D_1 and D_2 are the horizontal aggregation of the value of marginal product curves represented by equation (14.1). Labour supply is given by S_1 and S_2 respectively. In any period, profit maximisation implies equality of the wage and the value of the marginal product. Initial employment is E_1 and E_2 in regions 1 and 2, respectively. We are assuming that initially the real wage is w_0 everywhere, that is, let us initially consider a situation with equal amenities in both regions and a labour market that is in equilibrium. This is depicted in the top half of Figure 14.1 by the curves D_1, S_1, D_2 and S_2. Starting from this situation, region 1 benefits from a positive productivity shock per period, for example due to the greater scale of the agglomerated region yielding benefits from specialisation. This leads to a shift of region 1's demand curve to the right, to D_1' which puts upward pressure on the wage. As in standard labour market

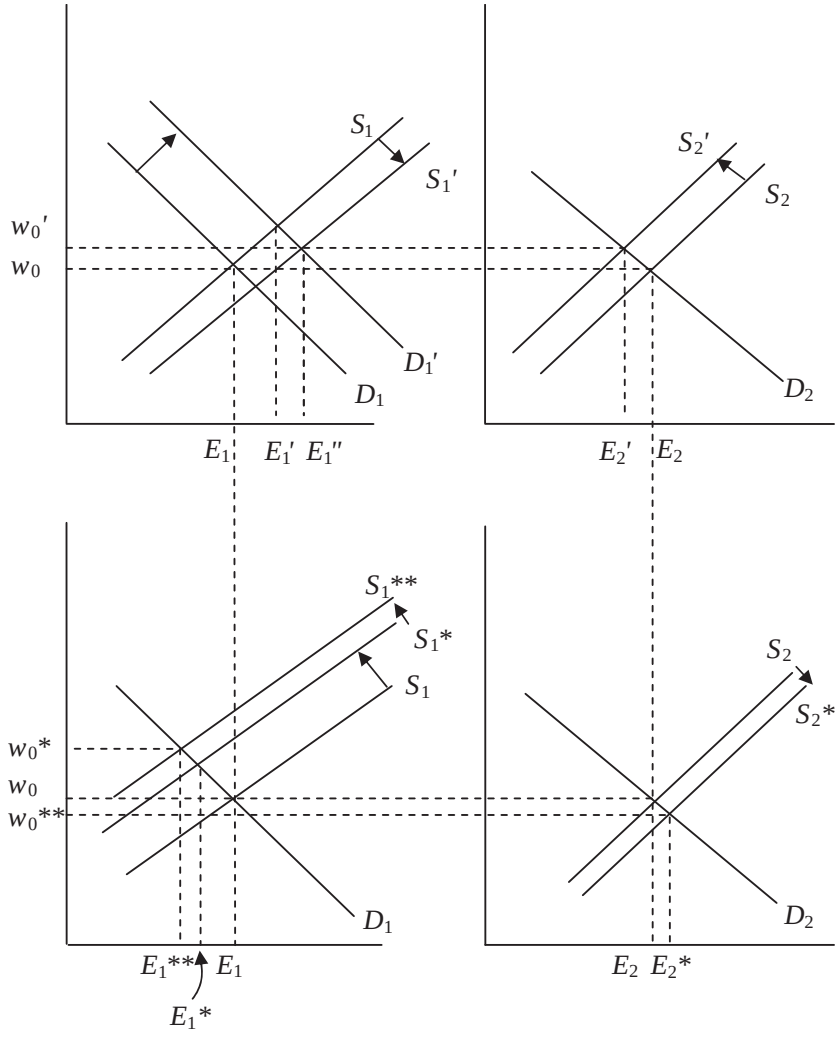


Figure 14.1 The dynamics of agglomeration externalities and interregional equilibrium

analysis, and assuming costless mobility, this generates increasing labour force participation, hours worked and inward migration that offsets some of the upward pressure on wages (top half of Figure 14.1). Net migration equals $E_1'' - E_1' = E_2 - E_2'$. In the new equilibrium, real wages are again equal and higher than initially (by $w_0' - w_0$) due to productivity having increased. Employment and the size of the economy of region 1 have increased while those of region 2 have decreased. It should be noted that this expansion of population and employment in region 1 may generate further dynamic externalities that may yield additional productivity growth and a further expansion of employment, that is, a virtuous circle of urban expansion.

However, this basic story can be complicated along various dimensions. Let us, for example, look at a situation in which the expansion of employment in region 1 on balance

has a negative utility externality effect on this region (we assume such effects are absent in region 2 for ease of exposition). The negative externality effect leads to a leftward, upward shift in the labour supply curve of region 1 (bottom left of the figure) to S_1^*, as workers demand a compensating differential. The vertical shift in the supply curve is equal to the size of this compensating (equilibrium) differential. This pushes up wages in region 1 to some extent, and will lead to some withdrawal of labour, but utility is subsequently nonetheless still higher in region 2. The consequence is outward migration from region 1 to region 2 and a shift in region 1's supply curve to S_1^{**}. In the new equilibrium, the wage in region 2 has declined somewhat from w_0 to w_0^{**} and the wage differential between the regions $w_0^* - w_0^{**}$ is exactly the compensating differential that leads to equal utility everywhere.

The combined effect of the positive and negative externalities (excluding further flow on effects of migration on productivity) in any given period is the sum of the shifts in the top half and bottom half of Figure 14.1. It can be seen that in the example there is overall an employment decline in region 2 (given by$(E_2 - E_2') - (E_2^* - E_2)$), while employment in region 1 is growing (given by $(E_1'' - E_1) - (E_1 - E_1^{**})$). Wages in the agglomerated region will increase by $w_0' - w_0 + (w_0^* - w_0)$, while those in the non-agglomerated region may increase or decline a little by $(w_0' - w_0) - (w_0 - w_1^{**})$ (since economy-wide total factor productivity growth is also not incorporated here).

In summary, we expect in the real world the positive effects in the agglomerated region to outweigh the negative effects on balance (as is consistent with the continued urbanisation observed worldwide). The combination of the effects is then likely to lead to both greater employment (due to the demand effect of the positive agglomeration externalities) as well as higher wages (to compensate for the negative externalities). It is the employment effect that is exploited in the empirical research by Glaeser et al. (1992).

The productivity shift on the right-hand side of equation (14.4) has one component that depends on time only. Neoclassical growth theory considers this to be the secular rate of technological change that applies throughout the economy and which is assumed constant over time and regions. Recently, however, there is increasing recognition that major innovations occur through the emergence of general purpose technologies at discrete, and unpredictable, points in time. Examples of these are the introduction of programmable computing networks in the twentieth century and of biotechnology and nanotechnology in the twenty-first century (Lipsey et al., 2005). More generally, innovation, technological change and the adaptation of workers and firms change productivity and equilibrium outcomes through equations (14.4) and (14.5) in complex ways that besides neoclassical modelling can also be analysed from evolutionary perspectives (for example Nelson and Winter, 2002).[6]

Given the model outlined above, the structure of the array $\mathbf{L}_t$ above provides proxies of measures that might be indicative of MAR, Porter and Jacobs externalities. This is the approach adopted by Glaeser et al. (1992) and several subsequent authors. The simplest measures of specialisation (S_{irt}), competition (C_{irt}) and diversity (D_{irt}) are respectively:[7]

$$S_{irt} = \frac{\sum_f L_{firt} / \sum_f \sum_i L_{firt}}{\sum_r \sum_f L_{firt} / \sum_r \sum_f \sum_i L_{firt}} \tag{14.6}$$

$$C_{irt} = \frac{F_{irt} / \sum_f L_{firt}}{\sum_r F_{irt} / \sum_r \sum_f L_{firt}} \tag{14.7}$$

$$D_{rt} = 1 - \sum_i \frac{\left(\sum_f L_{firt} \right)^2}{\left(\sum_j \sum_f L_{fjrt} \right)^2} \tag{14.8}$$

Equation (14.6) is just the definition of a location quotient, whereas equation (14.7) relates the inverse of firm size in a particular region and industry to the inverse of firm size in the national economy in that sector. Equation (14.8) is one minus the Herfindahl index of regional concentration of employment across sectors. In each region, this diversity measure is identical across industries. An industry-specific measure used by Glaeser et al. is the fraction of region *r*'s employment in the five largest industries other than industry *i* (measuring effectively a lack of diversity). A range of other, more advanced, measures is possible (see, for example, Maurel and Sedillot, 1999). It should also be noted that the measures above are essentially non-spatial (or, more precisely, topologically invariant) and that spatial interaction in the model is entirely by means of factor mobility (which is assumed to be costless).[8] Naturally, innovation diffusion is an explicitly spatial process that is not adequately captured in the simple measures above.

Glaeser et al. (1992) argue that the way in which the measures above affect employment growth depends on the type of externality considered. For example, under MAR externalities specialisation has a positive impact on productivity. Moreover, in these theories innovation is typically undertaken by large and dominant firms that can internalise the knowledge externalities. The impact of competition and diversity on growth would then be negative. In the context of Porter externalities, specialisation and competition are both positive forces, but diversity is not. Jacobs (1969) emphasised the importance of competition and diversity, while downplaying specialisation. These ideas are summarised in Table 14.1. The expected effects of localisation and urbanisation externalities (the latter including fiscal and environmental externalities) are also included in this table and are static in nature. Localisation externalities are not expected to create productivity growth in mature industries, but are at the heart of explanations for why cities exist in the first place and why they grew large in the past. This also holds for urbanisation externalities (including fiscal externalities) which typically have had a positive effect on employment, although they are increasingly dampened by congestion and pollution effects.

The theoretical literature has an empirical counterpart that aims at testing the hypotheses that are summarised in Table 14.1. This empirical literature strongly builds on the seminal contribution by Glaeser et al. (1992). In the next section, we provide a qualitative overview of this literature and the results obtained therein. Section 14.4 subsequently turns to a more in-depth description of the available empirical evidence on the various externalities and aims to provide an explanation for the variation in observed outcomes of the different studies.

Table 14.1 The effect of agglomeration externalities on employment

Type of externality			Effect on employment growth		
		Measured by	MAR	Porter	Jacobs
Dynamic	Knowledge externality	Specialisation	+	+	–
		Competition	–	+	+
		Diversity	–	–	+
Static	Localisation externality	Geography; Infrastructure		+	
	Urbanisation externality	Aggregate demand, metropolitan population		+	

14.3 A short review of recent empirical literature on agglomeration and growth

This section first discusses the way in which Glaeser et al. (1992) have simplified the model discussed in the previous section in order to arrive at a reduced-form equation that can be tested empirically. Next, we turn to a first description of the studies that were conducted following the seminal contribution by Glaeser et al. Apart from discussing the criteria that we adopted, including papers in the database underlying our meta-analysis, we also characterise those papers in terms of their outcomes, regional scope and the operationalisation of the dependent variable in the analysis (namely, urban growth).

The Glaeser approach

The study by Glaeser et al. (1992) builds on a very simple neoclassical model describing the functioning of an economy. The model can be seen as a simplified version of the general equilibrium model described in section 14.2. Central in their approach is a production function with 'technology' (A) and 'labour' (l) as inputs. Under perfect competition, profit-maximisation of individual firms results in equality of the marginal value product and the wage rate. Under the assumption of a simple Cobb–Douglas production function $y_{irt} = A_{irt} l_{irt}^{1-\alpha}$ (with i and r referring to industry and region respectively, as before), one arrives at the labour demand function:

$$l_{irt} = \left(\frac{\alpha A_{irt}}{w_{irt}} \right)^{\frac{1}{\alpha}} \tag{14.9}$$

Taking logs on both sides, one can easily arrive at an expression in growth rates:

$$\alpha \log\left(\frac{l_{irt+1}}{l_{irt}} \right) = \log\left(\frac{A_{irt+1}}{A_{irt}} \right) - \log\left(\frac{w_{irt+1}}{w_{irt}} \right) \tag{14.10}$$

This equation simply states that the growth rate of employment – *ceteris paribus* – positively depends on the growth of the state of technology, and negatively depends on the growth rate of wages. The growth rate of technology is subsequently assumed to depend on a national and a local component. The latter is explained from the three externalities identified in section 14.2, describing the impacts of specialisation, competition and diversity.

So we arrive at:

$$\log\left(\frac{A_{irt+1}}{A_{irt}}\right) = \log\left(\frac{A_{it+1,national}}{A_{it,national}}\right) + g(specialisation, competition, diversity) \quad (14.11)$$

which can be substituted into (14.10) to yield

$$\log\left(\frac{l_{ir,t+1}}{l_{irt}}\right) = -\frac{1}{\alpha}\log\left(\frac{w_{ir,t+1}}{w_{irt}}\right) + \frac{1}{\alpha}\log\left(\frac{A_{i,t+1,national}}{A_{i,t,national}}\right) + \frac{1}{\alpha}g(specialisation, competition, diversity)^{9} \quad (14.12)$$

The wage growth term is assumed to be the constant in regressions (that is, real wages grow equally across industries and regions) and changes in nationwide technology (and prices) are assumed to be captured by growth in nationwide industry employment. In order to test the empirical relevance of the various externalities, a dataset is constructed of growth rates of employment in a range of cities (MSAs) and mature industries.[10] These growth rates are subsequently regressed on a range of explanatory variables, among which the proxies for the three externalities are of key interest. Other explanatory variables are the aggregate growth of the industry considered at the national level, initial employment in the city-industry, and a dummy indicating presence in the south to allow for some sort of spatial heterogeneity. Overall, the results of the Glaeser study appear particularly consistent with the Jacobs perspective. The effect of specialisation as proxied by the location quotient of the city-industry is significantly negative. The effect of competition is positive, which is in line with the views expressed by Jacobs as well as Porter.

The study by Glaeser et al. (1992) was extended in a wide array of directions. It has been applied to different regions and different time periods, different proxies for the externalities have been used, growth has been operationalised in different ways, different estimation techniques have been used, and so on. Not surprisingly, these different approaches have led to different conclusions on the relevance of the various externalities in explaining growth. The aim of the remainder of this chapter is to provide an up-to-date account of the available studies and their results. Subsequently, we will try to get a grip on the sources of variation in the observed outcomes.

Selection and first characterisation of individual studies

In order to acquire a systematic and representative set of journal articles, we used Web of Science (www.isiknowledge.com) to select all articles that cited either Glaeser et al. (1992) or both Porter (1990) and Jacobs (1969). The result was a set of 318 articles covering the period up to April–May 2006. Our selection method results in a well-defined list, which is collected in a quick, efficient and reproducible manner. However, a consequence of this selection procedure is that it results in a list containing only journal articles. Mostly, no (as yet) unpublished articles, books or book chapters have been included. Furthermore, Web of Science has a bias towards journals written in the English language. To reduce the effects of the two negative effects associated with our selection method, we used the technique of snowballing, namely, carefully scanning through the references of the articles we included. This resulted in four more studies which Web of Science had not provided us with (among which one was French and one Italian).

We subsequently went through all the 322 articles, including in our database only those articles adopting a quantitative approach and including one or more variables

corresponding to any of the three variables for specialization, diversity and competition that Glaeser et al. (1992) introduced. In total, 31 articles were found to match Glaeser et al.'s methodology to a sufficient degree, giving us 393 different estimates.[11] They show considerable variation in the direction and significance of the effects found. Table 14.2 provides information on the studies included, the country to which the analysis pertains, the number of estimates provided by each study, and some characteristics of the dependent variable (namely, whether growth is defined in terms of employment, innovation, productivity, or otherwise). The table provides a first impression of the variation that is present in the studies. In the next section, we turn to a more elaborate statistical analysis of the available evidence.

14.4 Meta-analysis

Meta-analysis provides the researcher with a useful toolkit to study the sources of variation of study outcomes on particular topic. For excellent overviews of meta-analysis as a tool as well as for recent applications, see for example Florax et al. (2002) and Stanley (2001). This section will proceed by first summarising the available evidence by means of a simple vote count. Subsequently, we describe the results of our attempt to explain the observed variation in outcomes.

Vote counting

In order to get a first impression of the estimated effects of specialisation, competition and diversity, we have categorised all the available estimates into four classes: significantly negative, insignificantly negative, insignificantly positive and significantly positive. Ideally, we would have used a more refined effect size such as a (semi-)elasticity capturing the effects of specialisation, competition and growth. In the research under consideration, the heterogeneity in terms of both the dependent variable as well as the proxies used for our key variables of interest is so large that the construction of a common metric to characterise the available empirical evidence is not feasible (or, stated differently, leaves us with extremely small samples). As an aside, our approach implicitly builds on the assumption that all studies – regardless of the exact definition of their dependent variable – are informative on the determinants of growth. In other words, they require us to believe in a positive (possibly sequential) relationship between innovations,[12] patents, productivity and employment growth. For the moment, we will just make this assumption notwithstanding the fact that there is substantial theoretical literature on the relationships between growth, productivity, research and development (R&D), unemployment, and so on.[13]

The results of this vote-counting exercise are given in Table 14.3. Several results emerge. First, regarding specialisation there is no clear-cut evidence in the literature regarding its impact on the growth of cities. Although 70 per cent of the available estimates are statistically significant, of those about half are negative (the other half of course being positive). Regarding competition, results are somewhat clearer. Here 60 per cent of the estimated effect sizes are statistically significant and about two-thirds are positive, which is in line with Porter's hypothesis on the importance of competition in promoting urban growth. Finally, we consider the effects of diversity. Here only 50 per cent of the estimates are statistically significant. Out of those, however, more than 75 per cent point at a positive effect of diversity on urban growth.

Table 14.2 List of included studies and number of meta-observations

Study		Conclusions			Characteristics		
	# Est. eqs	SPEC	COMP	VARY	Country	Regions	Dependent
Sonobe and Otsuka (2006)	18	◦	n.a.	◦	Taiwan	Townships	9× empl., 9× other
Andersson et al. (2005)	12	n.a.	+	++	Sweden	LMAs	patents or innovations
Boschma and Weterings (2005)	5	◦	n.a.	−	Netherlands	NUTS3	patents or innovations
Acs and Armington (2004)	3	—	◦	n.a.	USA	LMAs	employment
Combes et al. (2004)	6	n.a.	◦	+	France	LMAs	other
Greunz (2004)	4	++	n.a.	++	Europe	NUTS2	patents or innovations
Lee et al. (2005)	5	—	++	++	South Korea	regions/counties	productivity
Malpezzi et al. (2004)	4	n.a.	n.a.	++	USA	SMAs	other
Mukkala (2004)	6	+	n.a.	n.a.	Finland	NUTS IV	productivity
Serrano and Cabrer (2004)	22	−	n.a.	◦	Spain	Provinces	productivity
Van der Panne (2004)	3	++	—	◦	Netherlands	ZIP regions	patents or innovations
Van Oort and Atzema (2004)	3	+	+	+	Netherlands	Municipalities	other
King et al. (2003)	7	−	++	◦	USA	States	employment
Rosenthal and Strange (2003)	18	+	◦	—	USA	ZIP regions	12× empl., 6× other
Batisse (2002)	6	—	◦	+	China	Provinces	other
Dekle (2002)	8	—	◦	◦	Japan	Prefectures	4× empl., 4× prod.
Massard and Riou (2002)	4	−	n.a.	−	France	Départements	patents or innovations
Staber (2001)	3	++	n.a.	—	Germany	circles of 10 km	Other
Combes (2000)	4	—	−	◦	France	LMAs	Employment
Baptista and Swann (1999)	4	+	◦	−	2× UK, 2× USA	CSO regions, states	Employment
Cainelli and Leoncini (1999)	4	++	++	++	Italy	Provinces	employment
Feldman and Audretsch (1999)	4	—	+	++	USA	SMAs	patents or innovations
Paci and Usai (1999)	6	++	n.a.	++	Italy	LMAs	patents or innovations
Partridge and Rickman (1999)	5	+	n.a.	+	USA	States	Productivity
Sjöholm (1999)	6	◦	◦	++	Indonesia	3× districts, 3× prov.	2×prod., 4× other
Baptista and Swann (1998)	9	−	n.a.	+	UK	CSO regions	patents or innovations
Bradley and Gans (1998)	1	n.a.	n.a.	—	Australia	Cities	Employment
Mody and Wang (1997)	6	—	+	n.a.	China	counties/provinces	productivity

Table 14.2 (continued)

Study	# Est. eqs	Conclusions			Characteristics		
		SPEC	COMP	VARY	Country	Regions	Dependent
Harrison et al. (1996)	7	∘	n.a.	n.a.	USA	Counties	patents or innovations
Henderson et al. (1995)	5	+	n.a.	∘	USA	SMAs	employment
Glaeser et al. (1992)	4	—	+	+	USA	SMAs	employment

Notes: The numbers in the second column indicate the number of estimated equations from which estimates for the externalities have been derived. The symbols in the next three columns have the following meaning: — significantly negative in all cases; – negative in all cases, but not always significantly so; ∘ inconclusive; + positive in all cases, but not always significantly so; ++ significantly positive in all cases; and n.a. no estimates available.

Table 14.3 Vote counts

	Specialisation		Competition		Diversity	
	Count	%	Count	%	Count	%
Negative significant	60	37	16	20	17	11
Negative insignificant	33	20	13	16	40	26
Positive insignificant	16	10	19	24	37	24
Positive significant	53	33	31	39	58	38
Total	162	100	79	100	152	100

Taken together, the first results of our meta-analysis tend to reconfirm the conclusions in Glaeser et al. (1992). There is substantial evidence for positive and significant effects of diversity and competition on urban growth, whereas the results regarding the effects of specialisation are highly ambiguous. In the next subsection we will provide a more detailed statistical analysis of the estimates that have been found in the literature and we will aim at explaining the sources of the variation that is present.

Meta-regression analysis

The previous discussions have pointed at the fact that both theoretically as well as empirically, there is lack of clear-cut evidence on the importance of the three dynamic externalities driving economic growth. This subsection aims to take the descriptive analysis in the previous section one step further by considering the relevance of several sources of heterogeneity. We proceed by first describing the potential sources of heterogeneity in study outcomes. Next, we describe the results of an ordered probit analysis and we conclude with a discussion of the main results.

Sources of variation in estimated effect sizes Some of the sources of variation were already identified in Table 14.2. They relate to the way in which the dependent variable in the analysis has been measured (namely, employment growth, productivity growth, patents or innovations, or other measures), the level of regional aggregation and the country covered in the analysis. Further heterogeneity is present in the sectoral coverage in the analysis. In our meta-analysis, we operationalise the characteristics of the dependent variable by means of several dummies and a continuous variable. The dummies measure whether the dependent variable is measured in terms of employment, patents or innovations, or productivity. Sectoral coverage is measured by two dummies that indicate whether the analysis is exclusively focused on the high-tech sector and whether the service sector has been included, respectively. Finally, we add a variable capturing the average population density of the units of observation included in the primary analysis. This captures in a simple and fairly comparable way an essential element of the regional aggregation of the analysis.[14]

A second set of factors that might affect the outcomes of the analyses concerns the empirical operationalisation of the key variables of interest, namely, specialisation, competition and diversity. First, the results for, for example, specialisation might be affected by the inclusion (or not) of a proxy for competition or diversity. Second, the exact empirical operationalisation can matter. Considering specialisation, it is likely to

matter whether specialisation is measured as a location quotient (namely, the share of a sector in regional employment relative to the national average) or just as a share in regional employment or total sectoral employment. For competition, different measures are used, among which number of establishments in a sector and the inverse of the average firm size in a sector feature most prominently. Regarding diversity, the crucial distinction is between studies that use the share of, for example, the five largest sectors and studies that use more continuous variables such as a relative diversity index, a Herfindahl index or a Gini coefficient. All these differences are captured by simple dummy variables.

A final set of factors that we consider relates to other data characteristics and the presence of additional control variables. These are: the period covered by the analysis (captured by the mean year of the analysis to which the data pertain); the length of the period covered (to distinguish between more long-run and short-run effects); the region covered in the analysis (taking Europe as the omitted category and considering Asia and the USA by means of dummies); the inclusion (or not) of investments, educational variables, wages and geographical variables as controls in the primary analysis; the estimation technique (distinguishing between panel and cross-sectional approaches); and the year of publication of the study.

Results from the ordered probit analysis In this section, we present the estimation results aimed at uncovering the factors explaining the direction and statistical significance of estimates obtained from the primary studies on the impact of specialisation, competition and diversity on urban growth. We estimate an ordered probit model distinguishing between the four ordered categories that were introduced in the section on vote counting. The estimation of an ordered probit model is common practice in a situation where the construction of a common metric to characterise the variation in the underlying primary studies is problematic. A downside of it is that it neglects information on the extent of statistical significance which is contained in, for example, the *t*-statistics of the estimated coefficients.[15]

The ordered probit model assumes the presence of a latent variable, y^* that can be explained by a set of explanatory variables x_i such that:

$$y^* = \sum_i \beta_i x_i + \varepsilon \tag{14.13}$$

where ε is an error term that is assumed to be well behaved. We only have information on the categorical variable y consisting of the four categories discussed above. This observed variable has the following structure:

$$\begin{aligned} y = 0 \quad & if \quad y^* \leq \mu_1 \\ y = 1 \quad & if \quad \mu_1 < y^* \leq \mu_2 \\ y = 2 \quad & if \quad \mu_2 < y^* \leq \mu_3 \\ y = 3 \quad & if \quad y^* > \mu_3 \end{aligned} \tag{14.14}$$

where the μ-parameters are estimated by the model, along with the β's. It is important to note that the interpretation of the estimated coefficients of an ordered probit analysis is

not straightforward, since the estimated coefficients only convey information on changes in the probability of finding an estimate in the extreme left and right category. In order to ease the interpretation of the results, we will focus our discussion on the results on the marginal effects which represent the change in the probably of finding an estimate in one of the four categories in response to a change of one of the explanatory variables.

The results of our ordered probit analysis are given in Table 14.4. The results for the variation in the effects of specialisation, competition and diversity on urban growth are given in the three respective columns. The explanatory variables capture the sources of variation that were discussed in the section on sources of variation in estimated effect sizes. For specialisation, competition and diversity, respectively, 60 per cent, 53 per cent and 59 per cent of the observations are predicted correctly by our model.

Before turning to a detailed discussion of the interpretation of these results, we compute the marginal effects. These facilitate the comparison of the outcomes for the different explanatory variables (see, for example, Greene, 2000, p. 878). The results are described in Table 14.5a–14.5c. All marginal effects are taken at the mean value of all other variables.[16] For the dummy variables, the marginal effects describe the increase in the probability of finding an outcome in one of the four categories of the dependent variable between the situation in which the value for a particular dummy is equal to zero and the situation in which it is one. For the continuous variables, the marginal effects are associated with an increase of the dependent variable by one. In case of the standardised variables, these correspond to an increase of the dependent variable by one standard deviation.

Discussion of the results In this subsection, we will discuss the most important results of our analysis as described in Tables 14.4 and 14.5. Let us first turn to the results regarding the characteristics of the dependent variable. For all three effects, the chances of finding significantly positive effects are substantially larger when measuring growth in terms of employment than in terms of productivity. This casts some doubts on the appropriateness of using employment as a proxy for technological development. Also interesting is that diversity tends to have a strongly significant positive effect if growth is measured in terms of innovation. This underlines the theory of Duranton and Puga (2001) who argue that innovation benefits from diversified or 'nursery' cities. Finally, it appears that the effect of diversity on urban growth is heterogeneous with respect to the sector considered. Studies that exclusively focus on the high-tech sector tend to find particularly strong and positive effects of diversity on urban growth. Carlino and Hunt (2007) find in this context that a more competitive local market structure increases innovation, but the effect is insignificant in Table 14.4 for studies of high-tech sectors.

Regarding the regions that are considered, we find that population density significantly and positively affects the chances of finding positive effects of specialisation on urban growth. This is an indication that indeed the level of spatial aggregation tends to matter for observed outcomes. Furthermore, the effects of specialisation, competition and diversity are hardly different between Europe and the USA. This result suggests that flexibility of goods and labour markets that differentiates – among many other factors – the USA from Europe has limited impact on the strength with which agglomerative forces function. These similarities are in contrast to Asia where the chances of finding positive effects for specialisation are limited, whereas the chances of finding positive effects for diversity are relatively large.

Table 14.4 Meta-regression analysis

	Specialisation	Competition	Diversity
Characteristics of dependent variable			
Data measure employment	0.54	0.41	1.26***
	(1.55)	(0.72)	(3.22)
Data measure patents or innovations	−0.24	−0.21	0.76*
	(–0.51)	(–0.26)	(1.97)
Data measure productivity	–0.97	−0.97	−0.88
	(−1.43)	(−0.92)	(−1.43)
Data are for high-tech only	−0.11	0.49	0.88***
	(−0.24)	(0.88)	(2.98)
Data include the service sector	0.03	−0.04	−0.06
	(0.23)	(−0.21)	(−0.65)
Specification of key variables			
Specialisation included		−1.87**	−0.70
		(−2.57)	(−1.42)
Specialisation as a location quotient	1.87***		
	(3.57)		
More specialisation variables included	0.01		
	(0.03)		
Competition included	−0.69		0.12
	(−1.14)		(0.24)
Competition is measured in est. per employee		0.99	
		(1.54)	
Competition is measured in establishments		1.57	
		(1.32)	
More competition variables included		−2.54**	
		(−2.20)	
Diversity included	0.71**	1.24*	
	(2.60)	(1.69)	
Diversity estimated using largest five			2.58***
			(3.34)
More diversity variables included			3.65***
			(6.23)
Other data characteristics			
Population density (log)	0.43***	−0.07	0.004
	(2.99)	(−0.21)	(0.03)
Standardised mean year to which the data pertains#	0.62**	0.42	0.92***
	(2.57)	(0.95)	(3.43)
Length of period covered by the data (in years)	0.74***	0.29	−0.01
	(3.19)	(0.69)	(−0.04)
Data are from Asia	−2.60***	0.06	1.88**
	(−3.41)	(0.06)	(2.47)
Data are from the USA	0.21	−0.33	−0.51
	(0.51)	(−0.39)	(−1.30)

Table 14.4 (continued)

	Specialisation	Competition	Diversity
Presence of additional control variables			
Investments or capital stock also included	2.31***	−0.57	−1.15
	(3.21)	(−0.38)	(−1.32)
Educational variables included	−1.99***	1.33**	2.36***
	(−4.95)	(1.99)	(3.75)
Wages or GDP also included	−0.51	−1.37*	0.001
	(−0.71)	(−1.96)	(0.00)
Geographical variables also included	−1.04**	−1.55	−0.29
	(−2.52)	(−1.63)	(−0.62)
Other study characteristics			
Estimated using panel data or similar	−1.31**	0.29	1.76**
	(−2.47)	(0.26)	(2.53)
Standardised year of publication#	0.32	−0.66	−0.17
	(1.36)	(−1.07)	(−0.72)
Limit point 1	−0.34	−1.03	−0.34
Limit point 2	0.49	−0.29	1.14
Limit point 3	0.89	0.57	2.49
Number of observations	162	79	152
Pseudo-R^2	0.26	0.22	0.40

Notes:
t-statistics are included in parentheses in the line below the estimate. Statistical significance is indicated with stars, where ***, ** and * reflect statistical significance at the 1, 5 and 10% level.
The variables are standardized in such a way that their mean is 0 and a value of +1 represents a value one standard deviation above the mean. For the mean year to which the data pertains, one standard deviation is 6.96; for the year of publication, it is 3.29.

A third set of results points at the potential importance of the time dimension. Both the effect of the length of the period covered in the analysis as well as the use of panel techniques (as opposed to pure cross-section techniques) are indicative in this respect. For specialisation in particular, it turns out that using cross-section techniques considering longer time periods tends to increase the chances of finding significantly positive effects. This can be interpreted as an indication that especially the effects of specialisation take time before they result in urban growth (using the fact that cross-section techniques and the consideration of long time periods help in identifying true long-run effects in primary analyses). We can also reason that apparently agglomeration forces still overcome negative externalities in the long run, and that therefore our findings support Glaeser's statement that cities are not dying (Glaeser, 1998).

A fourth set of results relate to the specification of the key variables of interest. Apart from the fact that the inclusion of specialisation, competition and diversity evidently have an impact on the estimated effects of the key variable of interest, two results stand out in particular. First, measuring specialisation as a location quotient (namely, relative to a national average) has a significantly positive effect on the chance of finding a positive effect of specialisation. This brings us to a more theoretical discussion as to whether it is

Table 14.5a Marginal effects: specialisation

	Neg. sign.	Neg. insign.	Pos. insign.	Pos. sign.
Data measure employment	−0.183*	−0.021	0.035*	0.169
	(−1.70)	(−0.65)	(1.75)	(1.43)
Data measure patents or innovations	0.090	−0.004	−0.020	−0.066
	(0.50)	(−0.22)	(−0.49)	(−0.54)
Data measure productivity	0.369	−0.070	−0.083	−0.215*
	(1.49)	(−0.77)	(−1.46)	(−1.94)
Data are for high-tech only	0.039	−0.001	−0.009	−0.029
	(0.24)	(−0.11)	(−0.24)	(−0.25)
Data include the service sector	−0.010	0.000	0.002	0.008
	(−0.23)	(0.02)	(0.23)	(0.23)
Competition included	0.256	−0.017	−0.056	−0.183
	(1.15)	(−0.47)	(−1.09)	(−1.26)
Diversity included	−0.272**	0.038	0.062**	0.172***
	(−2.57)	(1.05)	(2.11)	(3.10)
Specialisation as a location quotient	−0.510***	−0.141**	0.042	0.609***
	(−5.26)	(−2.25)	(1.38)	(3.91)
More specialisation variables included	−0.004	0.000	0.001	0.003
	(−0.03)	(0.00)	(0.03)	(0.03)
Population density (log)	−0.156***	0.000	0.034**	0.122***
	(−2.89)	(0.02)	(2.14)	(3.07)
Standardised mean year to which the data pertains	−0.225***	0.000	0.049**	0.176**
	(−2.65)	(0.02)	(2.05)	(2.45)
Length of period covered by the data (in years)	−0.271***	0.001	0.059**	0.212***
	(−3.24)	(0.02)	(2.35)	(2.96)
Data are from Asia	0.792***	−0.219***	−0.152***	−0.421***
	(7.16)	(−4.25)	(−4.10)	(−4.89)
Data are from the USA	−0.075	−0.002	0.016	0.061
	(−0.52)	(−0.21)	(0.52)	(0.5)
Investments or capital stock also included	−0.515***	−0.223***	−0.009	0.747***
	(−5.58)	(−3.50)	(−0.23)	(4.64)
Educational variables included	0.680***	−0.171***	−0.138***	−0.370***
	(7.18)	(−3.22)	(−3.73)	(−5.91)
Wages or GDP also included	0.198	−0.033	−0.046	−0.119
	(0.69)	(−0.36)	(−0.67)	(−0.93)
Geographical variables also included	0.391***	−0.064	−0.087**	−0.240***
	(2.62)	(−1.16)	(−2.36)	(−3.07)
Estimated using panel data or similar	0.485***	−0.157	−0.108***	−0.221***
	(3.02)	(−1.74)	(−2.76)	(−3.82)
Standardised year of publication	−0.117	0.000	0.025	0.091
	(−1.34)	(0.02)	(1.22)	(1.39)

Note: *t*-statistics are included in parentheses in the line below the estimate. Statistical significance is indicated with stars, where ***, ** and * reflect statistical significance at the 1, 5 and 10% level.

Table 14.5b Marginal effects: competition

	Neg. sign.	Neg. insign.	Pos. insign.	Pos. sign.
Data measure employment	−0.075	−0.068	−0.012	0.155
	(−0.77)	(−0.71)	(−0.36)	(0.71)
Data measure patents or innovations	0.045	0.034	−0.002	−0.076
	(0.24)	(0.27)	(−0.08)	(−0.27)
Data measure productivity	0.275	0.096***	−0.092	−0.280
	(0.72)	(2.76)	(−0.51)	(−1.32)
Data are for high-tech only	−0.079	−0.081	−0.027	0.187
	(−1.04)	(−0.88)	(−0.47)	(0.87)
Data include the service sector	0.008	0.007	0.000	−0.016
	(0.21)	(0.21)	(0.15)	(−0.21)
Specialisation included	0.223***	0.242***	0.184*	−0.650***
	(2.68)	(3.53)	(1.99)	(−3.64)
Diversity included	−0.337	−0.126***	0.096	0.366**
	(−1.32)	(−3.01)	(0.92)	(2.25)
Competition is measured in est. per empl.	−0.235	−0.129*	0.041	0.323*
	(−1.28)	(−1.98)	(0.56)	(1.85)
Competition is measured in establishments	−0.141**	−0.203**	−0.211	0.555*
	(−2.48)	(−2.28)	(−1.14)	(1.88)
More competition variables included	0.718**	0.074	−0.195***	−0.597***
	(2.42)	(0.65)	(−2.78)	(−3.76)
Population density (log)	0.014	0.012	0.001	−0.026
	(0.21)	(0.22)	(0.15)	(−0.21)
Standardised mean year to which the data pertains	−0.083	−0.069	−0.004	0.156
	(−0.95)	(−0.92)	(−0.22)	(0.97)
Length of period covered by the data (in years)	−0.057	−0.048	−0.003	0.108
	(−0.7)	(−0.67)	(−0.22)	(0.71)
Data are from Asia	−0.011	−0.009	−0.001	0.020
	(−0.06)	(−0.06)	(−0.04)	(0.06)
Data are from the USA	0.066	0.052	0.000	−0.118
	(0.37)	(0.41)	(0.03)	(−0.4)
Investments or capital stock also included	0.133	0.081	−0.022	−0.192
	(0.32)	(0.48)	(−0.17)	(−0.42)
Educational variables included	−0.161**	−0.194**	−0.140	0.495**
	(−2.39)	(−2.44)	(−1.34)	(2.33)
Wages or GDP also included	0.402	0.105**	−0.137	−0.370***
	(1.61)	(2.01)	(−1.22)	(−2.92)
Geographical variables also included	0.481	0.078	−0.179	−0.380***
	(1.39)	(0.82)	(−1.13)	(−3.12)
Estimated using panel data or similar	−0.049	−0.049	−0.014	0.112
	(−0.31)	(−0.25)	(−0.14)	(0.25)
Standardised year of publication	0.128	0.108	0.007	−0.243
	(1.02)	(1.06)	(0.21)	(−1.08)

Note: *t*-statistics are included in parentheses in the line below the estimate. Statistical significance is indicated with stars, where ***, ** and * reflect statistical significance at the 1, 5 and 10% level.

Table 14.5c Marginal effects: diversity

	Neg. sign.	Neg. insign.	Pos. insign.	Pos. sign.
Data measure employment	−0.013	−0.192***	−0.267***	0.471***
	(−1.61)	(−3.54)	(−2.83)	(3.72)
Data measure patents or innovations	−0.009	−0.134**	−0.153	0.295**
	(−1.42)	(−2.45)	(−1.55)	(2.01)
Data measure productivity	0.031	0.226	0.037	−0.293
	(0.74)	(1.27)	(0.54)	(−1.79)
Data are for high-tech only	−0.007	−0.133***	−0.120**	0.338***
	(−1.60)	(−3.37)	(−2.39)	(3.18)
Data include the service sector	0.001	0.012	0.009	−0.023
	(0.62)	(0.65)	(0.63)	(−0.65)
Specialisation included	0.007	0.117**	0.149	−0.273
	(1.54)	(2.03)	(1.10)	(−1.44)
Competition included	−0.002	−0.026	−0.020	0.048
	(−0.24)	(−0.24)	(−0.23)	(0.23)
Diversity estimated using largest five	−0.009	−0.184***	−0.476***	0.670***
	(−1.58)	(−4.42)	(−6.50)	(9.47)
More diversity variables included	−0.078**	−0.447***	−0.385***	0.909***
	(−2.19)	(−7.04)	(−6.51)	(19.87)
Population density (log)	−0.000	−0.001	−0.001	0.001
	(−0.03)	(−0.03)	(−0.03)	(0.03)
Standardised mean year to which the data pertains	−0.014*	−0.193***	−0.145**	0.352***
	(−1.71)	(−2.92)	(−2.53)	(3.55)
Length of period covered by the data (in years)	0.000	0.003	0.002	−0.005
	(0.04)	(0.04)	(0.04)	(−0.04)
Data are from Asia	−0.016*	−0.234***	−0.390***	0.639***
	(−1.86)	(−3.59)	(−2.97)	(3.65)
Data are from the USA	0.011	0.117	0.059	−0.186
	(0.87)	(1.22)	(1.47)	(−1.37)
Investments or capital stock also included	0.061	0.308	−0.031	−0.338**
	(0.64)	(1.32)	(−0.19)	(−1.98)
Educational variables included	−0.036*	−0.335***	−0.387***	0.757***
	(−1.66)	(−4.54)	(−4.97)	(6.33)
Wages or GDP also included	−0.000	−0.000	−0.000	0.000
	(0.00)	(0.00)	(0.00)	(0.00)
Geographical variables also included	0.005	0.064	0.037	−0.106
	(0.52)	(0.58)	(0.80)	(−0.65)
Estimated using panel data or similar	−0.008	−0.162***	−0.401***	0.571***
	(−1.57)	(−4.21)	(−3.52)	(5.06)
Standardised year of publication	0.003	0.037	0.027	−0.067
	(0.76)	(0.72)	(0.71)	(−0.73)

Note: *t*-statistics are included in parentheses in the line below the estimate. Statistical significance is indicated with stars, where ***, ** and * reflect statistical significance at the 1, 5 and 10% level.

absolute or relative size that matters in explaining variation in urban growth. It is not evident which is the preferable proxy for specialisation and scale. What is clear, however, is that the choice that is made tends to affect the outcome of the analysis substantially. Second, it stands out that studies that proxy diversity by means of a simple measure capturing the employment share of the five largest sectors (effectively measuring a lack of diversity) tend to find more positive effects of diversity than studies that use more refined measures to characterise diversity.

Finally, the inclusion of proxies for physical and human capital affect the outcomes for especially, specialisation and also diversity, whereas the inclusion of wages has a limited effect on the variation in outcomes in the primary studies. There also is no discernible effect of the year of publication.

14.5 Conclusions

This chapter has reviewed the theoretical background behind the empirical analysis of the growth of cities (and their hinterland) and subsequently looked into the available empirical evidence on the importance of three externalities in explaining urban growth, namely, MAR externalities, Porter externalities and Jacobs externalities. The latter was done by means of a meta-analysis. The overall evidence of the meta-analysis based on a simple counting of conclusive effect sizes reveals that relatively many primary studies conclude in favour of significantly positive effects of diversity and competition on growth. No clear-cut evidence was found for the effects of specialisation.

The meta-regression analysis points at several fruitful directions for further research. First, we found some quite strong indications for sectoral, temporal and spatial heterogeneity of the effects of specialisation, competition and diversity on urban growth. For example, high-tech sectors undergo stronger effects of diversity than other sectors, just as more recent data also find stronger effects for diversity and for specialisation. Finally, in studies using data from (East) Asia, specialisation is found to be a less important factor than elsewhere, while again diversity has more influence. Such heterogeneity typically remains unnoticed in primary studies which tend to focus the analysis on a specific region, sector or time period. It calls, for example, for research focusing on the dependency of the strength of agglomerative forces on the stage of development of the region, but also of the sector. This may enhance our insights into challenging questions as to whether in the knowledge-driven post-industrial economy of producer and consumer services characterised by many young and small firms, Jacobs externalities are more important. Second, the level of regional aggregation matters for the strength with which the agglomeration forces are operational. Especially for specialisation, population density has a positive influence on the results found. This gives rise to interesting questions regarding the transmission mechanisms through which the externalities function. More theoretical as well as empirical work investigating these issues is warranted. We also found that including control variables on investments or capital stock and education has substantial effects on our key variables of interest. Similar effects may be expected from factors such as social capital and trust, risk-taking and entrepreneurship, infrastructure, presence of multinational firms, R&D policies and institutions. More research on the role of such factors in determining the strength with which agglomerative forces are operating is warranted. Finally, we confirm the need for more attention to the specification of the key variables of interest. Again, further theoretical as well as empirical research along these lines is called for.

Notes

1. Key contributions in the economics literature aimed at understanding the growth of cities can be found in, for example, Acs (2006) and Black and Henderson (1999).
2. Roberts and Stanley (2005) provide a range of applications of meta-analysis in economics. See also Stanley (2001). A recent meta-analysis that complements the present chapter is Melo (2007), which focuses on the elasticity of output with respect to urban agglomeration. Melo finds an on average positive effect.
3. In our exposition, we abstain from an analysis of multi-region models in which development in a particular city also depends on developments in other (close-by) cities and factors such as the presence of multinationals through which regions may be connected to the global economy. Similarly, we do not pay explicit attention to, for example, the presence of infrastructure, labour competence, and so on (although they can easily be incorporated as elements of A in the production function). Nor do we account for the role of institutions (but see Henderson and Wang, 2007). Such factors are potentially relevant but remain underresearched in the empirical literature which we aim to survey in the remainder of this chapter. More attention in future empirical literature on such relationships is warranted.
4. So we assume, formally stated, that $w_{firt} = w_{irt}$ for each f. For simplicity, it is assumed that there is a one-to-one mapping between industries and occupations. Moreover, each industry produces only one commodity.
5. See, for example, Nijkamp and Poot (2004) for a meta-analysis of evidence of the impact of the macro level of education on growth.
6. See also Mulder et al. (2001) for a comparison of neoclassical, endogenous and evolutionary models of economic growth.
7. With respect to specialisation, some authors consider simplified relative measures such as $\frac{\sum_f L_{firt}}{\sum_r \sum_f L_{firt}}$, or even just absolute measures such as $\sum_f L_{firt}$.
8. See Duranton and Overman (2005) and de Dominicis et al. (2007) for studies that incorporate the spatial dimension more explicitly in an analysis of concentration.
9. Glaeser et al. (1992) actually allow the g function to vary also with initial conditions.
10. Only the six largest industries in each MSA are incorporated in the analysis.
11. These estimates were derived from 202 estimated equations, where most equations provided information on more than one externality. The number of estimated equations per study included in the database varies between 1 and 22 (see Table 14.2).
12. Arundel and Hollanders (2006) stress that the relationship between research and development (R&D), invention and innovation is a lot less clear than is often supposed among policy-makers. We could include R&D expenditure as an extra stage before innovations, using some kind of knowledge production function, but R&D was not to be found in the studies under consideration here (see Griliches, 1979, and Cameron, 1996, for an analysis of the effectiveness of R&D).
13. See, for example, Daveri and Tabellini (2000) and de Groot (2000) for some examples of studies in this area of research.
14. This variable describing the mean population density of the regions included in the study was constructed based on data on the regions included in the primary analyses (mainly from national statistical offices). We have also considered the average surface area and population size separately, but that did not lead to different results. Details are available upon request.
15. We refer to Koetse et al. (2006) for an example of an analysis along those lines and a comparison with a more simple ordered probit analysis.
16. Alternatively, we could have evaluated at the median. This turns out to have only limited impact on the outcomes. Details are available upon request.

References

Acs, Z.J. (2006), *The Growth of Cities*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.
Acs, Z.J. and C. Armington (2004), 'Employment growth and entrepreneurial activity in cities', *Regional Studies*, **38**, 911–27.
Andersson, R., J.M. Quigley and M. Wilhehnsson (2005), 'Agglomeration and the spatial distribution of creativity', *Papers in Regional Science*, **84**, 445–64.
Arundel, A. and H. Hollanders (2006), '*Searching the Forest for the Trees: "Missing" Indicators of Innovation*', Trend Chart Methodology Report, no. 2006, Maastricht.
Baptista, R. and G.M.P. Swann (1999), 'A comparison of clustering dynamics in the US and UK computer industries', *Journal of Evolutionary Economics*, **9**, 373–99.
Baptista, R. and P. Swann (1998), 'Do firms in clusters innovate more?', *Research Policy*, **27**, 525–40.
Barro, R.J. and X. Sala-i-Martin (2004), *Economic Growth*, Cambridge, MA: MIT Press.

Batisse, C. (2002), 'Dynamic externalities and local growth: a panel data analysis applied to Chinese provinces', *China Economic Review*, **13**, 231–51.
Black, D. and V. Henderson (1999), 'A theory of urban growth', *Journal of Political Economy*, 107, 252–84.
Bodvarsson, B., J.J. Lewer and H.F. Van den Berg (2007), 'Measuring immigration's effects on labor demand: a reexamination of the Mariel boatlift', unpublished manuscript.
Boschma, R.A. and A.B.R. Weterings (2005), 'The effect of regional differences on the performance of software firms in the Netherlands', *Journal of Economic Geography*, **5**, 567–88.
Bradley, R. and J.S. Gans (1998), 'Growth in Australian cities', *Economic Record*, **74**, 266–78.
Brakman, S. and B.J. Heijdra (2004), *The Monopolistic Competition Revolution in Retrospect*, Cambridge: Cambridge University Press.
Cainelli, G. and R. Leoncini (1999), 'Esternalità e Sviluppo Industriale di Lungo Periodo in Italia. Una Analisi a Livello Provinciale', *L'industria (NS)*, **20**, 147–66.
Cameron, G. (1996), 'Innovation and economic growth', PhD Thesis, University of Oxford.
Capello, R. (2004), 'Beyond optimal city size: theory and evidence reconsidered', in R. Capello and P. Nijkamp (eds), *Urban Dynamics and Growth: Advances in Urban Economics*, Amsterdam: Elsevier, pp. 57–85.
Carlino, G. and R. Hunt (2007), 'Innovation across US industries: the effects of local economic characteristics', Federal Reserve Bank of Philadelphia Working Paper 07-28.
Cheshire, P.C. and E.J. Malecki (2004), 'Growth, development, and innovation: a look backward and forward', *Papers in Regional Science*, **83**, 249–67.
Christaller, W. (1966), *The Central Places of Southern Germany*, Englewood Cliffs, NJ: Prentice Hall (first published in German in 1933).
Combes, P.P. (2000), 'Economic structure and local growth: France, 1984–1993', *Journal of Urban Economics*, **47**, 329–55.
Combes, P.P., T. Magnac and J.M. Robin (2004), 'The dynamics of local employment in France', *Journal of Urban Economics*, **56**, 217–43.
Daveri, F. and G. Tabellini (2000), 'Unemployment, growth and taxation in industrial countries', *Economic Policy*, **30**, 49–88.
de Dominicis, L., G. Arbia and H.L.F. de Groot (2007), 'Spatial distribution of economic activities in local labour market areas: the case of Italy', Tinbergen Institute Discussion Paper 07-094/3, Amsterdam-Rotterdam.
de Groot, H.L.F. (2000), *Growth, Unemployment and Deindustrialization*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.
Dekle, R. (2002), 'Industrial concentration and regional growth: evidence from the prefectures', *Review of Economics and Statistics*, **84**, 310–15.
Duranton, G. and H.G. Overman (2005), 'Testing for localization using micro-geographic data', *Review of Economic Studies*, **72**, 1077–1106.
Duranton, G. and D. Puga (2001), 'Nursery cities: urban diversity, process innovation, and the life cycle of products', *American Economic Review*, **91**, 1454–77.
Feldman, M.P. and D.B. Audretsch (1999), 'Innovation in cities: science-based diversity, specialization and localized competition', *European Economic Review*, **43**, 409–29.
Florax, R.J.G.M., H.L.F. de Groot and R.A. de Mooij (2002), 'Meta-analysis: a tool for upgrading inputs of macroeconomic policy models', Tinbergen Institute Discussion Paper, no. TI-2002-041/3, Amsterdam-Rotterdam.
Florida, R. (2002), *The Rise of the Creative Class: And How It's Transforming Work, Leisure, Community and Everyday Life*, New York: Basic Books.
Fujita, M. and J.F. Thisse (2002), *Economics of Agglomeration: Cities, Industrial Location, and Regional Growth*, Cambridge: Cambridge University Press.
Glaeser, E.L. (1998), 'Are cities dying?', *Journal of Economic Perspectives*, **12**, 139–60.
Glaeser, E.L. (2000), 'The new economics of urban and regional growth', in G.L. Clark, M.P. Feldman and M.S. Gertler (eds), *The Oxford Handbook of Economic Geography*, Oxford: Oxford University Press, pp. 83–98.
Glaeser, E.L., H.D. Kallal, J.A. Scheinkman and A. Shleifer (1992), 'Growth in cities', *Journal of Political Economy*, **100**, 1126–52.
Greene, W.H. (2000), *Econometric Analysis*, Upper Saddle River, NJ: Prentice-Hall.
Greunz, L. (2004), 'Industrial structure and innovation: evidence from European regions', *Journal of Evolutionary Economics*, **14**, 563–92.
Griliches, Z. (1979), 'Issues in assessing the contribution of research and development to productivity growth', *Bell Journal of Economics*, **10**, 92–116.
Harrison, B., M.R. Kelley and J. Gant (1996), 'Specialization versus diversity in local economies: the implications for innovative private-sector behavior', *Cityscape: A Journal of Policy Development and Research*, **2**, 61–93.
Henderson, J.V. and H.G. Wang (2007), 'Urbanization and city growth: the role of institutions', *Regional Science and Urban Economics*, **37**, 283–313.

Henderson, V., A. Kuncoro and M. Turner (1995), 'Industrial development in cities', *Journal of Political Economy*, **103**, 1067–90.
Jacobs, J. (1969), *The Economy of Cities*, New York: Random House.
King, C., A.J. Silk and N. Ketelhohn (2003), 'Knowledge spillovers and growth in the disagglomeration of the US advertising-agency industry', *Journal of Economic Management Strategy*, **12**, 327–62.
Koetse, M.J., H.L.F. de Groot and R.J.G.M. Florax (2006), 'The Impact of Uncertainty on Investment: A Meta-analysis', Tinbergen Institute Discussion Paper, no. TI 2006-060/3, Amsterdam-Rotterdam.
Krugman, P.R. (1995), *Development, Geography, and Economic Theory*, Cambridge, MA: MIT Press.
Lee, B.S., K. Sosin and S.H. Hong (2005), 'Sectoral manufacturing productivity growth in Korean regions', *Urban Studies*, **42**, 1201–19.
Lipsey, R.G., K.I. Carlaw and C.T. Bekar (2005), *Economic Transformations: General Purpose Technologies and Long Term Economic Growth*, Oxford: Oxford University Press.
Lösch, A. (1954), *The Economics of Location,* New Haven, CN: Yale University Press (first published in German in 1940).
Malpezzi, S., K.Y. Seah and J.D. Shilling (2004), 'Is it what we do or how we do it? New evidence on agglomeration economies and metropolitan growth', *Real Estate Economics*, **32**, 265–95.
Marshall, A. (1890), *Principles of Economics: An Introductory Volume*, London: Macmillan.
Massard, N. and S. Riou (2002), 'L'Impact des Structures Locales sur l'Innovation en France: Spécialisation ou Diversité?', *Revue Région et Développement*, **16**, 111–36.
Maurel, F. and B. Sedillot (1999), 'A measure of the geographic concentration in French manufacturing industries', *Regional Science and Urban Economics*, **29**, 575–604.
Melo, P. (2007), 'Urban agglomeration and productivity: a meta-analysis', Centre for Transport Studies, Imperial College London.
Mody, A. and F.Y. Wang (1997), 'Explaining industrial growth in coastal China: economic reforms . . . and what else?', *World Bank Economic Review*, **11**, 293–325.
Mukkala, K. (2004), 'Agglomeration economies in the Finnish manufacturing sector', *Applied Economics*, **36**, 2419–27.
Mulder, P., H.L.F. de Groot and M.W. Hofkes (2001), 'Economic growth and technological change: a comparison of insights from a neoclassical and an evolutionary perspective', *Technological Forecasting and Social Change*, **68**, 151–71.
Nelson, R.R. and S.G. Winter (2002), 'Evolutionary theorizing in economics', *Journal of Economic Perspectives*, **16**, 23–46.
Nijkamp, P. and J. Poot (1998), 'Spatial perspectives on new theories of economic growth', *Annals of Regional Science*, **32**, 7–37.
Nijkamp, P. and J. Poot (2004), 'Meta-analysis of the effect of fiscal policies on long-run growth', *European Journal of Political Economy*, **20**, 91–124.
OECD (2005), *Oslo Manual: Guidelines for Collecting and Intepreting Innovation Data*, OECD: Paris.
Ohlin, B. (1933), *Interregional and International Trade*, Cambridge, MA: Harvard University Press.
Paci, R. and S. Usai (1999), 'Externalities, knowledge spillovers and the spatial distribution of innovation', *GeoJournal*, **49**, 381–90.
Partridge, M.D. and D.S. Rickman (1999), 'Static and dynamic externalities, industry composition, and state labor productivity: a panel study of states', *Southern Economic Journal*, **66**, 319–35.
Porter, M.E. (1990), *The Competitive Advantage of Nations*, Houndmills: Macmillan.
Roberts, C.J. and T.D. Stanley (2005), *Meta-regression Analysis. Issues of Publication Bias in Economics*, Oxford: Blackwell Publishing.
Rosenthal, S.S. and W.C. Strange (2003), 'Geography, industrial organization, and agglomeration', *Review of Economics and Statistics*, **85**, 377–93.
Serrano, G. and B. Cabrer (2004), 'The effect of knowledge spillovers on productivity growth inequalities in Spanish regions', *Environment and Planning A*, **36**, 731–53.
Sjöholm, F. (1999), 'Productivity growth in Indonesia: the role of regional characteristics and direct foreign investment', *Economic Development and Cultural Change*, **47**, 559–84.
Solow, R.M. (1956), 'A contribution to the theory of economic growth', *Quarterly Journal of Economics*, **70**(1), 65–84.
Sonobe, T. and K. Otsuka (2006), 'The division of labor and the formation of industrial clusters in Taiwan', *Review of Development Economics*, **10**, 71–86.
Staber, U. (2001), 'Spatial proximity and firm survival in a declining industrial district: the case of knitwear firms in Baden-Wurttemberg', *Regional Studies*, **35**, 329–41.
Stanley, T.D. (2001), 'Wheat from chaff: meta-analyis as quantitative literature review', *Journal of Economic Perspectives*, **15**, 131–50.
Swan, T.W. (1956), 'Economic growth and capital accumulation', *Economic Record*, **32**, 334–61.
UNFPA (2007), *State of World Population 2007: Unleashing the Potential of Urban Growth*, New York: UNFPA.

Van der Panne, G. (2004), 'Agglomeration externalities: Marshall versus Jacobs', *Journal of Evolutionary Economics*, **14**, 593–604.
Van Oort, F.G. and O.A.L.C. Atzema (2004), 'On the conceptualization of agglomeration economies: the case of new firm formation in the Dutch ICT sector', *Annals of Regional Science*, **38**, 263–90.
Zhang, J. (2003), 'Growing Silicon Valley on a landscape: an agent-based approach to high-tech industrial clusters', *Journal of Evolutionary Economics*, **13**, 529–48.

15 Sustainable development and regional growth

Amitrajeet A. Batabyal[1] *and Peter Nijkamp*

15.1 Introduction

Regional development does not take place in a wonderland of no physical-geographical dimensions. The region is sometimes an abstract concept, but sometimes also a very concrete, real-world geographic space where actual economic forces are manifesting themselves. The realization of economic growth is conditioned by constraints and opportunities emerging from the environmental and resource base of a region. A balanced regional growth perspective calls for a thorough investigation of environmental, resource and climatological conditions that are responsible for sustainable development.

Two trends in recent social science research have increasingly guided the nature of contemporary research in regional science. As noted by Batabyal and Nijkamp (2004), the first is the recognition by regional scientists that many outstanding problems in regional science have a distinct environmental dimension to them. The second is the acknowledgment by natural resource and environmental economists – see Stevens and Olsen (2004) – that effective renewable resource management and environmental externality regulation cannot be divorced from considerations of the space over which the management and the regulatory functions are to be undertaken. These two trends together have now given rise to a rather substantial literature on topics at the interface of regional science and the environment.

Even though there is no gainsaying the existence of this sizeable literature on regional science and the environment, this literature is widely scattered over a large number of books and journals and, to the best of our knowledge, there are virtually no syntheses of the principal themes in this burgeoning literature. Therefore, our basic objective in this chapter is to review the key themes in this sizeable literature. To this end, in the rest of this chapter, we shall endeavor to be adequately broad and deep at the same time. The reader should note that our review is both retrospective and forward-looking. We discuss what has been accomplished thus far and the likely future directions of research on regional science and the environment.

We have structured our primary arguments in this chapter as follows. We first contend that the availability of resources and environmental quality are intimately connected with industrial and service development in a particular area and with the welfare of individuals living in cities and regions in this area. Second, we point out that regional economic development involves the utilization of scarce natural resources. In this regard, it is salient to grasp that the natural environment provides a useful resource-base that can support economic growth. Even so, the degradation of this resource-base can deleteriously impact upon both environmental quality and economic growth. Third, we remind the reader that spatial environmental externalities and non-market transactions will often provide a rationale for regulatory interventions by governments. Fourth, we note that at any pertinent spatial scale of analysis, there is typically a need for a clear mapping of the spatial implications of environmental degradation. Such mappings will frequently lead to the use

of apposite tools such as geographic information systems that can act as useful empirical inputs in spatial-economic and environmental models. Finally, we discuss the significant regional responses that will be required to mitigate and adapt to the ongoing effects of climate change.

We now narrow our discussion and concentrate on five key themes in the extant research on regional science and the environment. We do this in large part because we believe that the most significant and thought-provoking new research in regional science is likely to involve one or more of these five themes. These five themes are: (1) regional economic development; (2) natural resources; (3) environmental regulation; (4) geographic information systems; and (5) regional climate change. The remainder of this chapter is organized as follows. Section 15.2 focuses on environmental issues in the context of regional economic development. Section 15.3 discusses issues pertaining to natural resource use in a regional setting. Next, section 15.4 concentrates on the nature of regulatory policy when environmental and regional issues are pertinent Then, section 15.5 focuses on the present and the possible future uses of geographic information systems (GIS). Section 15.6 addresses the repercussions of and the responses to regional climate change and variability. Finally, section 15.7 offers concluding comments.

15.2 Regional economic development

Researchers having even a nodding acquaintance with the regional science literature know that the subject of regional economic development has fascinated scholars for many decades. Even so, it is only fairly recently that researchers have begun to investigate the occasionally broad and sometimes narrow environmental dimensions of regional economic development. In this regard, attention has been focused sometimes on narrow growth issues such as air pollution caused by an industrial plant or the noise caused by airplanes, and at other times on broad issues such as the effect of industrial transformations and biodiversity loss in a given area. Therefore, we now comment on three features of regional economic development in which the environment, either directly or indirectly, is a substantive factor that has received a lot of attention. These features are the local industrial structure, the integration of ecology in regional analysis, and the local institutional environment.

Industrial location and structure

The location and the structure of economic units has a salient bearing on whether regional economic development is or is not sustainable. Therefore, one can ask what effect the local economic environment has on business location. Gabe (2007) has focused on the state of Maine in the United States to answer this question. He shows that although businesses are attracted to areas with high short-run and seasonal stability, with the exception of service businesses, annual fluctuations in local employment do not deter business activity. Hence, policy-makers can 'jointly pursue the objectives of local economic growth and stability' (Gabe, 2007, p. 398).

The role of location in affecting alternate aspects of regional economic development has been studied by researchers in other settings as well. For instance, Gress and Poon (2007) note that although inter-firm relations do affect the location and the investment decisions of Korean firms in the United States, what they call intra-firm and extra-firm relations have an even more salient impact on these decisions. Similarly, using empirical

analysis, Portnov et al. (2007) have argued that in Nepal, location is a salient determinant of urban economic growth. This is because the fastest-growing urban localities are all close to major population centers, to highways, and to the border with India.

Just as industrial location has a bearing on regional economic development, so does industrial structure. The structure of economic units certainly has significant effects on the daily quality of life. Air pollution, noise, water pollution and many other kinds of emissions lead to unpriced impacts, very often at a local or at a regional level. To comprehend variation in the local quality of life, one needs to have insights into the reasons for regional decline and industrial clustering. In this regard, Polese and Shearmur (2006) use examples from Canada to argue that regional decline will become an increasingly common occurrence in nations that are at the end of the demographic transition and whose economic geographies display what these researchers call 'center–periphery' relationships. The phenomenon of industrial clustering has been interestingly studied by Bottazzi et al. (2007). These researchers show that agglomeration is typically the outcome of technology-specific drivers and site-specific geographical forces. The environment, broadly construed, is clearly an important determinant of the rate and the extent of regional economic development. As such, we now proceed to shed light on this particular issue in contemporary research in regional science.

Environmental quality

The state of the physical environment, measured by a set of apposite indicators, determines the multifaceted notion known as environmental quality. In other words, environmental quality refers to the condition of the natural or the built environment at a point in time (space) or over time (space). In regional environmental policy studies there is the usual trade-off between economic development and environmental quality. Even so, despite the importance of regions, the connection between environmental quality and economic activity and growth has typically been studied in a non-regional setting. In an attempt to correct this lacuna in the literature, Van den Bergh and Nijkamp (1998) – also see Van den Bergh and Nijkamp (1991, 1997) and Verhoef et al. (1997) – have analyzed a theoretical model of multi-regional growth, environmental processes and multi-regional trade. In the non-coordination model, each region attempts to grow optimally, taking the actions of the second region as given. In this setting, Van den Bergh and Nijkamp (1998) show that free-rider benefits over time may result. In contrast, in the coordination model, the rate of economic growth in each region will be slower (faster), as environmental externalities dominate (are dominated by) technological externalities.

What impacts do public environmental policies have on the concentration of firms engaging in risky environmental activities? Using a monopolistic competition framework, Calmette and Pechoux (2006) show that contrary to normal expectations, environmental policies that affect the marginal costs of firms exacerbate the concentration of firms engaging in risky activities. In contrast, when one puts in place what these authors call 'regionally differentiated regulation', one is able to limit effectively the agglomeration of firms whose activities are likely to cause environmental accidents.

Clearly, environmental quality issues are consequential not only in theoretical settings but in practical settings as well. Here, the work of Barajas et al. (2007) and Smith (2007) is germane. Barajas et al. (2007) have analyzed the environmental performance of foreign companies operating in the region demarcated by the northern border of Mexico with the

United States. These authors show that because of cooperation between extant companies and the government, the region under study has not become a repository of 'dirty' industries. Similarly, Smith (2007) has used data to study the relevance of 'race'-versus 'economic deprivation'-based explanations of urban inequality in Detroit, Michigan. His results indicate that economic deprivation supercedes race as an explanatory factor and that when attempting to comprehend the phenomenon of environmental inequality, it is necessary to recognize the decisive role played by the process of deindustrialization.

Local institutional environment

Several scholars have pointed out that what drives regional economic change is the presence and the significance of local institutional networks.[2] For instance, Mitchell-Weaver (1992) – along with Piore and Sabel (1984) and Scott and Storper (1986) – has used a broad interpretation of the institutional environment to understand the nature of regional economic change in Pittsburgh in the United States and in the surrounding regions. As he points out, the local environment in the Pittsburgh area is dominated by the presence of two vital institutional networks: the integrated regional industrial complex and the public–private partnerships built up through the Allegheny Conference on Community Development (ACCD). Does it make sense to appeal to the policies of these two networks to explain the empirically observed changes in the economy of the Pittsburgh area? Mitchell-Weaver's (1992) analysis does not lead to an unambiguous conclusion. He concludes that: 'much more work is needed to affirm the effectiveness of regional innovation networks and to evaluate in what circumstances and in which time frames they might work' (Mitchell-Weaver, 1992, p. 285).

With an eye on the local institutional environment, Lambert and Boerner (1997) have addressed the contentious subject of environmental justice. A salient question in this context is the following: are economic factors the primary cause of environmental inequity? Yes, say Lambert and Boerner (1997). Consequently if regional economic development is to be 'environmentally just', then it will be necessary to institute 'a policy that compensates individuals living near industrial sites' (Lambert and Boerner, 1997, p. 195). Clearly, such a perspective also calls for the use of neo-Rawlsian principles in environmental policy.

A broad view of the institutional environment can be seen in the work of researchers such as Partridge (2007) and Press (2007). Focusing specifically on rural economic development, Partridge (2007) has noted that the combination of what he calls 'urban-centered rural growth' and higher energy costs suggests that policy-makers ought to pay attention to the following two relevant findings. First, more regional planning mechanisms will be needed to ensure the participation of rural areas in urban growth. Second, the deleterious effects of high energy costs in remote rural areas will require special attention to support infrastructure investments and rural to urban in-migration.

What role do institutional arrangements play in firm cluster adjustment as a result of a changing economic environment? Press (2007) has shed useful light on this question. She focuses on two scenarios in which firms either act in their own interest (the egoistic case) or in their joint interest (the collective case). Her analysis shows that when the underlying institutional arrangements are unstable, the collective outcome is unlikely to emerge because firms will typically not act in their joint interest. In contrast, when the same institutional arrangements are stable, firm clusters are more likely to demonstrate what she

calls 'collective local cultures'. As our discussion thus far in this sections has shown, regional development studies of all sorts focus either directly or indirectly on the environment. In addition, the recent discussion on sustainability has given rise to a great deal of interest in the intricate connections between regional economic developments and environmental quality indicators.

Outlook

A significant amount of research has now been conducted on environmental issues in the context of regional economic development. This notwithstanding, there certainly are a number of outstanding research questions in this area that need to be studied. In this regard, the reader should note that in the next several years, the number of private and public regional development policies will generally increase. Further, these policies will, most likely, change environmental quality in the pertinent region(s). Therefore, there is a great need to develop metrics that can be used to measure the benefits and the costs of regional development policies that have environmental ramifications. In this regard, Morisugi and Ohno (1995) propose to use a so-called 'benefit incidence matrix' to evaluate the benefits and the costs of the above kinds of policies.[3] Similarly, Wen et al. (2007) have suggested that researchers use the genuine progress indicator (GPI) to measure regional economic performance in general and urban economic welfare in particular. Although the matrix approach and the GPI approach are certainly useful, a lot more work needs to be done in order to comprehensively measure the benefits and the costs of regional development policies with environmental implications.

In recent times, the notion of 'regional sustainability' has increasingly become a popular concept. Several researchers such as Giaoutzi and Nijkamp (1994), Gutman (2007) and Wallis et al. (2007) have written about this concept. Although most authors thus far have referred to sustainability in the sense of Brundtland (1987), a lot of new and interdisciplinary research on this concept is still needed. In particular, because measurement criteria for judging sustainable development are unclear, the following sorts of queries need to be researched. First, what information is needed to meet the challenge of Brundtland sustainability in a regional setting? Second, if a macroeconomic accounts approach is deficient in dealing with the problem of regional sustainability, then how should one design green regional product accounts that pay sufficient attention to local and to regional informational needs? Third, is there a trade-off between regional economic growth and regional environmental quality? Finally, what is the optimal degree of interdependence between regionally sustainable economies? These are some of the main questions that await further study by researchers. These questions are challenging to study because, *inter alia*, the inclusion of environmental, ecological and social factors necessarily involves both locational specificity and difficulties in comparability.

15.3 Natural resources

Natural resources exist in regions and hence we can think of regions as the geographic base for a wide variety of natural resources. In addition, renewable natural resources such as forests and water provide humans with an assortment of vital consumptive and non-consumptive services. The utilization of natural and other resources often leads to the emission of substances that eventually have a detrimental effect on the earth's climate. As well, the use of natural and other resources produces waste. Consequently, researchers in

a variety of disciplines have analyzed questions relating to the optimal use of natural resources and waste management. Given this situation, we now comment on the more important themes concerning natural resource use in regional settings. We do this by focusing on three specific topics, and these topics are deforestation, water provision and waste management.

Deforestation

Several issues concerning forests have occupied the research attention of regional scientists. This notwithstanding, as Nelson et al. (2004) and Walker (2004) have pointed out, one theme that has increasingly come to dominate much of the present center stage in discussions of forests is the subject of deforestation. Commercial logging of forests, particularly in tropical areas, is carried out mainly by means of concession logging. Hence, if we are to understand the phenomenon of tropical deforestation, then we must first comprehend the institution of concession logging and the ways in which appositely designed forest management principles can prevent unsustainable concession logging practices.

In concession logging, a key management question is the following: how does a regulator ensure that loggers will not cut trees in excess of allowed limits within the relevant contract period? Walker and Smith (1993) have analyzed this question rigorously. Their study leads to two noteworthy policy conclusions. First, these researchers show that partial inspection policies designed to discover non-compliance early in a contractual period can be very cost-effective. Second, we learn that 'longer contract lengths tend both to increase renewal values of future contracts and to increase the effectiveness of partial inspection policies' (Walker and Smith, 1993, p. 415).

With deforestation in the background, researchers such as Nelson et al. (2004) and Pfaff et al. (2007) have asked: relative to new road construction, what is the role of road improvements in causing deforestation? Using data for Darien province in Panama, Nelson et al. (2004) have shown that whereas both new road construction and road improvements lower transport costs, relative to the original road development, this cost reduction is a lot less. Therefore, it is important to comprehend location-specific impacts, and these researchers argue that to comprehend properly these location-specific effects, one needs to use either the nested multinomial logit model or the random parameters logit model. Pfaff et al. (2007) have found evidence of spatial spillovers from roads in the Brazilian Amazon. Specifically, they show that deforestation rises in census tracts that lack roads but are in the same county as and within 100 kilometers of a tract with a new paved or unpaved road.

Water provision and use

Water is one of the main sources of human well-being. Regional authorities all over the world are frequently given the task of providing citizens with vital natural resources such as water. Therefore, researchers have devoted considerable attention to studying the optimal ways of providing and using water in nations as diverse as Bangladesh (Akmam and Higano, 2007), Ghana (Hunter, 2006) and the United States (Troesken, 2002). Do regional authorities provide citizens with water cheaply? Can we tell whether present water provision practices are efficient? Is decentralization in the provision of water across regions a good idea? These are some of the important questions that researchers have studied thus far.

Akmam and Higano (2007) begin their interesting study of safe water provision in Bangladesh by pointing out that approximately 80 million people in Bangladesh are presently in danger of losing their health and achieving a much lower life expectancy because of exposure to arsenic contamination in their drinking water obtained from tube wells. They then use a multi-objective mixed-integer programming model to evaluate the various factors impinging on the decision to provide safe water to the affected Bangladeshis. Their analysis shows that the 'type of safe water' that is optimal depends on the underlying simulation case being analyzed.

Bhattacharyya et al. (1995) have examined the provision of water in the rural regions of the state of Nevada in the United States. They note that because of the presence of a number of unfunded state and federal mandates, cost control has become a primary goal of rural water utilities. Specifically, these utilities would like to know whether they are providing water to their constituents efficiently, that is, as cheaply as possible. To answer this empirical question, Bhattacharyya et al. (1995) use a hedonic shadow cost function approach and show that the Nevada utilities under study are allocatively inefficient because they are using excessive amounts of energy relative to labor. Even so, because these water utilities face constraints that are generally not considered in neoclassical efficiency analyses, the policy implications of the above allocative inefficiency result are unclear. In fact, it is noted that it is possible for 'the allocative performance [of these utilities] . . . to get worse although the utilities' managerial performance may not be worsened' (Bhattacharyya et al., 1995, p. 498).

In many parts of the Intermountain West in the United States, on account of rapid population growth and rising water development costs, governments have attempted to condition residential development approval on the adequacy of water supplies. What impacts have such regulations had on housing supplies? This question has been ably addressed by Hanak and Chen (2007) with data for the states of Colorado and New Mexico. Using fixed-effects panel regressions, these researchers have shown that relative to quantity controls, price-based regulatory tools designed to ensure water availability are preferable. In addition, these researchers have also shown that attempts to 'restrict groundwater basin access have not unambiguously corrected negative externalities related to growth' (Hanak and Chen, 2007, p. 85). These questions of water scarcity and water provision play an important role not only in developed countries but also in developing countries where, in addition to the usual economic considerations, cultural and gender-related factors often play a salient role in the collection and the distribution of water.[4]

Waste management

Waste is a significant by-product in any industrial nation. Although it is a source of environmental stress, it may also become an input for recycling activities. Plainly, there are many locational issues involved. For instance, where should we put our garbage? Ye and Yezer (1997) use a theoretical model of optimal noxious facility siting to analyze this important question. The basic idea in this paper is to compare the efficient facility siting equilibrium with a collective choice equilibrium in which voters recognize the benefits and the costs of alternate garbage disposal site location patterns. The authors show that collective choice outcomes can be remarkably inefficient and that 'policies adopted under majority rule voting result in overly large facilities at which there is underexpenditure on pollution control compared to an optimal solution' (Ye and Yezer, 1997, pp. 65–6).

Given the above finding, is it a better idea to collect garbage conventionally and then use landfills, or does it make more sense to recycle garbage? McDavid (2000, p. 157) has used survey data and has shown that even in situations that are favorable to recycling, 'recycling is more expensive than conventional collection and landfilling'. A similar point has been made by Kinnaman (2006) in his study of the 8875 municipalities in the United States that had initiated curbside recycling programs over the previous two decades to help reduce residential solid waste. Using the empirical lessons learned in the previous two decades from solid waste management in the United States, Kinnaman (2006) has argued for the replacement of several state recycling mandates with a reasonable landfill tax.

Outlook

Natural resource use in a regional setting is a vast subject. As such, many significant research questions presently remain unanswered. As far as forest use is concerned, it would be useful to identify empirical methodologies that will allow a researcher to study the impact of road improvements on deforestation as part of a more complete evaluation of the expansion of infrastructure for economic development. On a separate note, research is needed to differentiate clearly between regional water-providing utilities that are allocatively efficient and those that are managerially efficient. In addition, it would be helpful to identify policies that inefficient utilities can use to become efficient.

As far as waste management is concerned, the generality of the extant literature's findings with regard to the desirability of landfilling over recycling needs to be explored. Specifically, it would be useful to identify general conditions under which either landfilling or recycling will lead to a minimization of the total cost of waste management. Finally, given the recent study by Sicotte and Swanson (2007), additional research on the location of landfills and other waste facilities and apposite governmental policies is needed to ensure that minorities and the poor do not always end up living near such facilities.

15.4 Environmental regulation

As a result of burgeoning interest in studying the formulation and the implementation of environmental policy – see, for example, Cohen (2006) – there now exists a sizeable literature on this subject. This literature has concentrated on several important questions and therefore this section is devoted to a discussion of three of these salient questions.[5]

Effect on regional economic activity

Do environmental regulations cost jobs? Cole and Elliott (2007) have shed light on this important question by focusing on the case of the United Kingdom. They first focus on the case where environmental regulations are exogenous and they then allow environmental regulations and employment to be determined endogenously. Their analysis shows that irrespective of whether environmental regulation costs are exogenous or endogenous, there is no statistically significant impact of these regulations on employment. Put differently, there is no evidence of a trade-off between jobs and the environment.

Similar to the above question, we can ask what the relationship is between per-unit pollution abatement compliance costs and regional economic activity. Duffy-Deno (1992) has used a sample of 63 Standard Metropolitan Statistical Areas (SMSAs) to conduct a detailed empirical analysis of this salient question. Using econometric analysis, Duffy-Deno (1992,

p. 419) concludes that there is 'weak support for the argument that environmental regulations retard economic activity'.

Along the same lines, Garofalo and Malhotra (1995) have used a model based on James Tobin's (1969) q theory of investment to analyze the effect of environmental regulations on regional capital formation at the state level. The main goal of their paper is to explore the following idea: if environmental regulations raise a firm's cost of production, then this should reduce a firm's q value[6] and therefore reduce its rate of capital formation. Garofalo and Malhotra (1995) use panel data on 34 states and seven time periods and find that the 'effect of environmental regulations on net capital formation is modest' (p. 214).

The results from the previous three paragraphs suggest that environmental regulations have little or no impact on regional economic activity. However, the work of Lee (2007) clearly shows that this is not always the case. Focusing on Korean manufacturing industries, Lee (2007, p. 91) points out that 'environmental regulations caused a 12 percent decline in the average annual rate of productivity growth over the period 1982–93'. More generally, Millimet and List (2004) have used what they call a 'method of matching' to demonstrate that the effect of strict environmental regulations is heterogeneous across space and that this effect varies systematically with location-specific characteristics. Therefore, studies that assume a homogeneous response to environmental regulations across space are likely unintentionally to mask the overall effect of more stringent environmental regulations by pooling affected and unaffected regions. This discussion tells us that there is now a clear need for more reliable, meta-analytic research in this area.

Acceptability of Pigouvian taxes

On more occasions than has typically been recognized, there is a distinct spatial dimension to external diseconomies such as pollution. Consequently, it is certainly reasonable to ask whether it is adequate to use Pigouvian taxes to secure an optimal allocation of resources in the presence of pollution. Further, when the pertinent external diseconomies are transboundary in nature, how might we analyze the spatial aspects of corrective environmental policies? These sorts of questions have interested researchers in both environmental economics and regional science.

In a prominent book, Baumol and Oates (1988, p. 54) have contended that: 'economic efficiency requires the absence of compensation of victims of detrimental externalities . . . in the case where the affected entities are relatively small'. This standpoint has been contested by a number of researchers such as Carlton and Loury (1980, 1986). Lately, Uimonen (2001) has used a spatial general equilibrium model and noted that a single instrument such as a Pigouvian tax is not adequate to restore economic efficiency in a polluted environment. This is because in a competitive setting with pollution, the same environment is both a common property resource for firms entering the industry and an external diseconomy for present victims. Put differently, there are two distortions in the economic environment. Hence, in general, the use of a single instrument such as a Pigouvian tax will not lead to an optimal allocation of resources.

Using a two-region spatial price equilibrium model, Verhoef and Nijkamp (2000) have analyzed the conduct of environmental policy in a setting in which the relevant external diseconomies are transboundary in nature. In their model, trade, environmental spillovers and uncoordinated taxes result in interactions between the two regions that are being studied. These researchers study the relative merits of consumption taxes,

production taxes and a hybrid instrument that is part consumption tax and part production tax.

A case for the use of suitably adjusted Pigouvian taxes has been made by Verhoef (2003) in his study of the problems posed by traffic congestion. Using a continuous-time and continuous-place dynamic model of traffic congestion, Verhoef (2003) shows that when departure times are endogenous and there is a bottleneck along the relevant route, 'hypercongestion' arises on the upstream road segment in an intertemporal equilibrium. To deal with this phenomenon, Verhoef (2003, p. 531) advocates the use of an instrument such as a congestion toll because congestion tolls 'based on an intuitive dynamic and space-varying generalization of the standard Pigouvian tax rule can hardly be improved upon'. Verhoef's advocacy notwithstanding, it must be said that more applied and real-world-based research is needed to determine the extent to which Verhoef's (2003) results are general and the usefulness of alternate corrective policy instruments.

Purpose of regionalization

Solutions to the problem of external diseconomies have a clear geographic dimension to them. Therefore, it is pertinent to focus on contemporary models of new economic geography and to ask how these models have dealt with external diseconomies. In this regard, the recent paper by Hosoe and Naito (2006) is representative. Building on the prominent Krugman (1991) model, these authors have studied the nature of regional agglomeration effects in a two-region model in which there is transboundary pollution. Specifically, these authors have analyzed the impacts of environmental damage and the subsequent environmental tax on the distribution of the population between the two regions. The analysis undertaken shows that the equilibrium pattern of population distribution is the same in the short run in which households cannot migrate from one region to the other, and in the long run in which this kind of migration is possible.

A model of interregional household mobility has also been used by Wellisch and Richter (1995), but to shed light on issues pertaining to the local control of stock pollutants and the public debt. The analysis by Wellisch and Richter (1995) shows that the local public debt is generally not neutral. In addition, assume that the implicit factor rewards to local pollution are left with landowners to avoid the migration distortions of mobile households. In this case, there is a lot of scope for regional internalization because household mobility obligates local authorities to account for the marginal willingness to pay of all future generations living in the region, when controlling the present pollution level.

This positive role of regionalization raises a related question. When there is imperfect information about one or more aspects of pollution control, how ought central and local authorities to interact? Andersen and Jensen (2003) answer this question by noting that if the central authority is highly uncertain about the environmental effects of a specific pollutant then a tax or subsidy scheme ought to be designed to permit local information to play a key role in the conduct of environmental policy. In contrast, if the central authority is certain that a pollutant must not exceed a particular limit then a similar tax or subsidy scheme ought to be designed to permit local information little influence in the conduct of environmental policy. A similar question, but now with the additional twist of corruption, has been analyzed by Hu et al. (2004). These authors show that in a society with no bribery, a higher local fine share or higher fines will clearly increase pollution

abatement. In contrast, in a corrupt society, a higher local fine share or higher fines may reduce pollution abatement.

Outlook

We saw in the section on the effect on regional economic activity that an inability to recognize the fact that the impact of exacting environmental regulations is spatially heterogeneous has led several studies to assume a spatially homogeneous response to environmental regulations and hence these studies have unwittingly hidden the overall effect of more stringent environmental regulations by pooling affected and unaffected regions. Therefore, there is ample scope for methodological research in this area.

With regard to the acceptability of Pigouvian taxes, it would be useful to gain additional knowledge about the attributes of environmental taxation and other regulatory instruments such as zoning in a spatial setting in which households are heterogeneous. Also, following the work of Quaas (2007), it would be worthwhile to develop and analyze policy instruments that will enable us to solve urban environmental problems efficaciously. *Inter alia*, this will enable us to study the nexuses between population growth in cities and the related task of increased infrastructure provision.

Hosoe and Naito (2006) have already shown us that the equilibrium pattern of population distribution is the same in the short run in which households cannot migrate from one region to the other, and in the long run in which this kind of migration is possible. However, this result is true in a two-region model. Consequently, additional research is needed to ascertain the generality of this result in multi-region models. In addition, in pollution control and other settings in which imperfect or incomplete information is an issue, new research is needed to determine the optimal interaction between central and local regulatory authorities.

15.5 Geographic information systems

At various points in this chapter, we have underscored the need for appropriate quantitative research. However, in order to conduct appropriate quantitative research, we must first have the right information on region–environment interactions. In this context, advances in area-specific information systems are noteworthy. A geographic information system (GIS) is a computer system capable of assembling, storing, manipulating and displaying geographically referenced information, that is, data identified according to their location. GIS technologies can be used for a variety of purposes. Increasingly, they are being used in agricultural economics,[7] for regional planning, and for natural resource and environmental management.[8] Accordingly, in the rest of this section, we shall first discuss the use of GISs in two pertinent contexts: residential issues and spatial-environmental data. Then, we shall contend that GIS technologies offer a number of new opportunities for raising our comprehension of disaggregate human spatial behavior.

Residential issues

What can the use of a GIS tell us about the accessibility of housing to public community facilities (PCFs)? This question has been taken up by Shen (2002). Using a GIS for four county-wide metropolitan areas in North Carolina in the United States, Shen (2002, p. 235) shows that 'distinct housing accessibility patterns exist'. What this means is that multi-family housing such as apartments and townhouses enjoy higher accessibility to

desirable PCFs. In contrast, what Shen (2002) calls 'manufactured housing' is typically farthest from desirable PCFs and actually closer to some undesirable PCFs.

The decision to buy or sell a house is clearly related to the spatial characteristics of the property itself. This notwithstanding, when studying the buy–sell decision, most researchers thus far have overlooked the spatial aspect of the underlying story. Therefore, Kim and Horner (2003) have attempted to fill this gap in the literature by incorporating exogenous spatial variables into their Cox proportional hazards model by using GIS-based modeling techniques. They show that spatial factors (commuting time zones) and non-spatial factors (equity constraints) are both salient in explaining housing turnover.

The use of GIS modeling techniques has been shown to be useful in the case of residential location decisions as well. Arguing that the determinants of home and workplace location choices depend on an individual's life cycle, Kim et al. (2005) have examined the home–workplace location choice decision from a commuting standpoint. Their GIS-based analysis shows that environmental amenities and attributes are often very useful in explaining the nexus between commuting behavior and residential location choices. This kind of work is useful not only because it demonstrates the use of a GIS, but also because it shows that when a GIS is complemented with other tools, a researcher's ability to conduct meaningful spatial analysis is enhanced significantly.

Spatial data and environmental analyses

Researchers now agree that location is an important factor in determining both land use and land values. Traditionally, scholars have dealt with the idea of location by appealing to unidimensional measures such as access or distance. However, this approach is problematic because the significance of location in influencing land use and land values is not limited to accessibility alone. Indeed, as Geoghegan et al. (1997, p. 252) have noted, externalities 'characterize land use, and these externalities are spatially determined'. In addition, the measurement of the amenity values that are connected to either a specific landscape pattern or to a mosaic of natural and human-managed patches is expedited by examining GIS data.

Today there is no controversy on the point that GISs provide us with a technology that can be used effectively to tackle environmental problems by integrating spatially consistent data from a number of different sources (Campbell and Masser, 1995). In addition, the scope for meaningful analysis using GISs is very broad. Therefore, in the remainder of this section, we focus on three studies that demonstrate the diversity of questions that can be addressed using GIS modeling techniques.

How sustainable are residential blocks of differing physical densities? Ghosh et al. (2006) have shed light on this question by focusing on the case of five residential blocks in Auckland, New Zealand. These authors pay attention to five dimensions of sustainability and they then use aerial photographs, GISs and ecological footprint assessment techniques to compute domestic energy demand, generation and waste. The analysis undertaken in this paper shows that the New Zealand suburb in which the physical density is 18 households per hectare has the greatest potential to be sustainable.

How does one undertake integrated spatial assessment (ISA) comprehensively? Girard and De Toro (2007) focus on a rural village in southern Italy – San Marco dei Cavoti – and argue that comprehensive ISA requires the effective integration of what they call an analytic hierarchy process (AHP) and a GIS. According to these researchers, such an

integration will enable planners to determine effectively whether the cultural and the environmental heritage of a village such as San Marco dei Cavoti is or is not being developed in a sustainable manner.

Finally, how can a GIS be used to determine the flood mitigation benefits of wetlands? Ming et al. (2007) answer this question by concentrating on the Momoge National Nature Reserve in Jilin province in the People's Republic of China. These researchers use a GIS to first estimate the flood mitigation capacity of wetland soils in the above reserve. Next, they convert this capacity into a monetary measure of the economic benefits from flood mitigation. On the basis of this analysis, Ming et al. (2007) convincingly claim that the quantitative analysis of flood mitigation benefits will be a worthwhile reference for the assessment of wetland values in the reserve and for more general discussions about wetland functions and their usefulness.

Outlook

In the foregoing two sections, we first discussed the way in which a GIS can be used to analyze residential issues and then we commented on some ways in which environmental analyses can be made more beneficial to society by carefully combining a GIS with other methodologies. These uses notwithstanding, the analysis of human spatial behavior is a quickly growing one and a GIS environment is an ideal one in which to analyze many of the outstanding research questions. For instance, Longley (1998) has noted that researchers are now able to access various kinds of digital spatial data that they could not access even a few years ago. Because these data sets contain detailed information not available before, they will permit researchers to shed light on hitherto unexamined questions about many aspects of the urban environment such as the usefulness of alternate street planning procedures and transportation networks.

Anselin (2000) has noted that the toolbox of spatial econometrics in particular and the study of space in general will need to be extended to address acceptably the challenges posed by the analysis of socio-economic, space–time data. Particular issues that will require future attention include the estimation of space–time dynamics for limited dependent data, the modeling of changing choice sets, and the design of techniques to distinguish between spatial dependence and spatial heterogeneity.

Environmental monitoring and natural resource management are labor-intensive and costly to undertake. Therefore, it is important to look for ways to reduce the high labor requirements and the associated costs. Tsou (2004) has shown that this can be done by integrating Web-based GISs with image-processing tools. Specifically, he has pointed out that such integration can provide easy access to geospatial information and to Web-based image analysis. In turn, this can be used to alter the detection capabilities of natural resource managers and regional park rangers. This is a very useful beginning and more work on these sorts of topics will bring the full scope of GIS-based research to the forefront.

15.6 Regional climate change

A variety of gases such as carbon dioxide, methane, nitrous oxide and water vapor are known as greenhouse gases because much like the glass in a greenhouse, they trap infrared radiation that would normally escape into the earth's atmosphere. This entrapment tends to have a warming effect on the globe and this 'global warming' can ultimately result in

climate change. There is almost no debate on the proposition that the greater the level of greenhouse gases, the greater the equilibrium temperature of the earth. There is also very little debate on whether anthropogenic emissions of greenhouse gases cause a significant rise in global temperature in comparison with current temperature levels and in comparison with natural fluctuations in the temperature levels. As noted by Kahn (1998, p. 167), the 'debate centers around the magnitude and timing of the change, and its significance to human welfare'.

There now exists a vast literature that has studied the problem of climate change at the global level. However, the work of Ruth (2006a), Ruth et al. (2006), Smith and Mendelsohn (2006) and Calzadilla et al. (2007) clearly tells us that there is a distinct regional dimension to the problem of global climate change and, in addition, global climate change has stochastic regional impacts. Hence, the study of regional climate change and its impacts is salient in its own right. Accordingly, in the rest of this section, we discuss two important ways in which regional climate change alters our thinking about, first, ecosystem management, and second, the nexuses between local and global climatic conditions.

Ecosystem management

The work of Gleick (1987), Burn (1994) and Ruth et al. (2006) tells us that higher temperatures can lead and have led to changes in snowfall and in snowmelt dynamics in specific regions such as mountainous watersheds in California. The ensuing loss of snow and ice cover combined with a rise in ocean temperatures and the thermal expansion of the water mass in the oceans has resulted in an increase in the average global sea level of 0.1 to 0.2 meters in the twentieth century. In addition, the impacts of changes in ocean temperatures, sea levels and coastal storm patterns have had dramatic impacts on ecosystems in general. Although the impact on some ecosystems such as fisheries may well be positive (see below), in many instances wetlands have been lost, shorelines have been eroded, groundwater has become salinized, and the ecological and the economic values of protected areas have been diminished.

Consider the case of protected areas in Africa. Velarde et al. (2005) first classify different African ecosystems in accordance with the so-called Holdridge Life Zone (HLZ) system. They then use a benefits transfer approach to place an economic value on the predicted ecosystem shifts resulting from climate change in the protected areas under study. They point out that there are 20 HLZs in Africa and all of these HLZs can be found in the protected area network. The analysis undertaken by these researchers suggests that urgent action by the managers of these protected areas is needed to arrest the decline in economic values. This is because for the year 2100, in three of the four global circulation models, there is a negative economic impact of climate change in the protected areas of Africa. In addition, we learn that in certain alternate model scenarios, the total economic damage up to 2100 can be as large as US$74.5 million.

Expected climate change in the twenty-first century is likely to increase ocean temperatures in the North Atlantic and this expected increase is likely to have an impact on important fish stocks in the Icelandic–Greenland ecosystem and hence on the economy of this region. However, will these anticipated impacts be positive or negative? Arnason (2007) has attempted to answer this question by providing estimates of the impact of altered fish stocks due to global warming on the Icelandic and Greenland economies. He

undertakes several stochastic simulations and points out that the impact of global warming on the Icelandic gross domestic product (GDP) is more likely to be positive than to be negative. For Greenland, the effect of global warming on fish stocks and the GDP is very likely to be positive and substantial relative to the current GDP. These findings point to two important general lessons for ecosystem management. First, the effects of regional climate change are not necessarily negative. Depending on the location of a particular ecosystem, the effect of global warming on ecological and economic variables may well be positive. Second, there is still a great deal of uncertainty about the magnitude of both positive and negative effects associated with regional climate change. This requires ecosystem managers to be cautious, particularly when their actions may result in either hysteresis or irreversibilities.[9]

Nexuses between local and global climatic conditions

Urban and other land use changes often account for as much as half the observed increase in the diurnal temperature range in certain regions (Ruth et al., 2006). Road construction materials such as concrete, buildings and other structures, tend to absorb and not reflect the sun's heat. In addition, the removal of trees and shrubs eliminates the natural cooling effects of shading and evapo-transpiration. These factors together frequently give rise to a regional warming phenomenon known as the urban 'heat island effect'. This effect is different from global warming but the effect itself may influence and be influenced by global climate change. It is important to comprehend that urban heat island effects correspond to localized shifts in climate that approximate changes in global climate projected to occur over the next 100 years. Therefore, by studying how cities impact and are impacted upon by climate change, 'we may begin to formulate strategies for mitigating and adapting to rising urban temperatures attributable to both regional and global warming phenomena' (Stone, 2006, p. 318). Given this state of affairs, we now discuss two studies of urban heat island effects in two very different parts of the world.

Focusing on the city of Phoenix in the United States, Guhathakurta and Gober (2007) ask what aspects of smart growth policies planners ought to concentrate on when there are urban heat island effects to contend with. They undertake statistical analysis and their analysis shows that increasing daily low temperatures by 1° Fahrenheit is associated with a mean monthly increase in water use of 290 gallons for a typical single family unit. This and other similar results tell us that when attempting to create more compact urban forms, planners ought to consider carefully the effects of heat island effects on the demand for water. In addition, they also ought to account for the other environmental consequences of heat island effects in their evaluation of smart growth strategies.

The Pearl River Delta (PRD) in the southern part of Guangdong province in the People's Republic of China is one of the world's most rapidly developing regions. What are the impacts of rapid urbanization and its associated heat island effects on the local and the regional atmospheric circulations over the PRD? Lo et al. (2007) have shed useful light on this interesting question. Their analysis using numerical experiments leads to two noteworthy conclusions for planners. First, it is shown that stronger urban heat island effects in the PRD will increase the differential temperature gradient between urbanized areas and the nearby ocean surface. Second, these researchers point out that further industrial development and urbanization will increase the air temperature in the lowest 2 kilometers of the atmosphere.

Outlook

The discourse in this sixth section tells us that although the problem of global climate change is an important one, there are several meaningful issues at a sub-global scale that warrant serious study. Two such issues concern ecosystem management in the presence of regional climatic variations, and the nexuses between local and global climatic change. More generally, following the discussion in Ruth (2006b) and in Ruth et al. (2006), one needs to recognize that the trinity of land use decisions, regional development and environmental quality are very closely related. Further, this relationship itself is multidimensional in nature. Therefore, additional research is needed to comprehend not only the complexities in the relationship between the above-mentioned trinity, but also the ways in which this complexity manifests itself over time – from the short run to the long run – and over space – from local to regional and global scales.

Ecosystem management in the presence of regional climate change involves decision-making under fundamental uncertainty. Not only are researchers uncertain about the values of key model parameters but they are also uncertain about the full impacts of alternate managerial actions. In such settings, the use of the prominent precautionary principle[10] is likely to be useful and a lot will be gained by keeping two points in mind when contemplating the task of ecosystem management. First, ecosystems such as fisheries and forests are really ecological–economic systems whose dynamic and stochastic behavior is determined by forces that are partly ecological and partly economic in nature (Batabyal, 1999). Second, as much of the contemporary literature in ecological economics has shown,[11] in order to optimally manage an ecological–economic system, it will be important to focus on a concept such as the Holling resilience of this system.[12]

15.7 Conclusions

In this chapter, we have conducted a retrospective and forward-looking review of the sizeable literature on issues concerning regional development and the environment. Specifically, we discussed the many ways in which the environment has entered and influenced research pertaining to: (1) regional economic development; (2) natural resources; (3) environmental regulation; (4) geographic information systems; and (5) regional climate change. Our review demonstrates that the environment is definitely here to stay in regional science research. As we have pointed out in the various 'Outlook' sections, even though several noteworthy research questions have now been adequately studied by scholars, there are a number of outstanding research questions that have received little or no attention from them. Therefore, in the years to come, one may look forward to many new and interesting developments concerning these hitherto unexplored research questions. Addressing these new challenging issues is definitely a *sine qua non* for sustainable regional development.

Notes

1. Batabyal acknowledges financial support from the Gosnell endowment at RIT. The usual disclaimer applies.
2. For more details on this issue in a Dutch land use context, see Nijkamp et al. (2002).
3. There are many ways to conduct project analysis. One way is to use a benefit incidence matrix (BIM). As Morisugi and Ohno (1995) have pointed out, this technique is useful when a project gives rise to environmental effects and these effects need to be accounted for. A benefit incidence matrix is typically based on a socio-economic model which itself is often constructed using multi-regional general equilibrium theory. When the main focus of an analyst is on tracking the sectoral impact of changes in policy on an entire

region's economy, a social accounting matrix (SAM) is helpful. As Berck and Hoffman (2002) have noted, the SAM framework is an extension of the more traditional technique of input–output analysis.
4. See Hunter (2006) for a more detailed corroboration of this claim.
5. This literature has an international dimension as well. For more on this, see Batabyal and Nijkamp (2004).
6. The q value of a firm is the ratio of the firm's market value to its replacement cost. See Tobin (1969) for more details.
7. For more details, see the various papers in Nelson (2002).
8. As an example, a GIS might permit emergency planners easily to compute emergency response times in the event of a natural disaster. Alternately, a GIS might be used to locate wetlands that need protection from pollution.
9. For more on hysteresis and irreversibilities, see Barham et al. (1998), Islam et al. (2003) and Wirl (2006).
10. See Silva and Jenkins-Smith (2007) for more on the precautionary principle.
11. See Batabyal and Beladi (1999), Batabyal and Godfrey (2002) and Batabyal and Yoo (2007a, 2007b) for more details on this literature.
12. See Batabyal and Yoo (2007a, 2007b) for more on the concept of an ecological–economic system's Holling resilience.

References

Akmam, W. and Y. Higano (2007), 'Supplying safe water in Bangladesh: a policy model based on multi-objective mixed integer programming', *Papers in Regional Science*, **86**, 57–75.

Andersen, P. and F. Jensen (2003), 'Local pollution in federal systems', *Environmental and Resource Economics*, **26**, 417–28.

Anselin, L. (2000), 'GIS, spatial econometrics, and social science research', *Journal of Geographical Systems*, **2**, 11–15.

Arnason, R. (2007), 'Climate change and fisheries: assessing the economic impact in Iceland and Greenland', *Natural Resource Modeling*, **20**, 163–97.

Barajas E., M.D., C. Rodriguez C. and H. Garcia J. (2007), 'Environmental performance of the assembly plants industry in the north of Mexico', *Policy Studies Journal*, **35**, 265–89.

Barham, B.L., J. Chavas and O.T. Coomes (1998), 'Sunk costs and the natural resource extraction sector: analytical models and historical examples of hysteresis and strategic behavior in the Americas', *Land Economics*, **74**, 429–48.

Batabyal, A.A. (1999), 'Contemporary research in ecological economics: five outstanding issues', *International Journal of Ecology and Environmental Sciences*, **25**, 143–54.

Batabyal, A.A. and H. Beladi (1999), 'The stability of stochastic systems: the case of persistence and resilience', *Mathematical and Computer Modelling*, **30**, 27–34.

Batabyal, A.A. and E.B. Godfrey (2002), 'Rangeland management under uncertainty: a conceptual approach', *Journal of Range Management*, **55**, 12–15.

Batabyal, A.A. and P. Nijkamp (2004), 'The environment in regional science: an eclectic review', *Papers in Regional Science*, **83**, 291–316.

Batabyal, A.A. and S.J. Yoo (2007a), 'A probabilistic approach to optimal orchard management', *Ecological Economics*, **60**, 483–86.

Batabyal, A.A. and S.J. Yoo (2007b), 'A stochastic analysis of the Holling resilience of an orchard', *Ecological Economics*, **64**, 1–4.

Baumol, W.J. and W.E. Oates (1988), *The Theory of Environmental Policy*, 2nd edn, Cambridge: Cambridge University Press.

Berck, P. and S. Hoffman (2002), 'Assessing the employment impacts of environmental and natural resource policy', *Environmental and Resource Economics*, **22**, 133–56.

Bergh, J.C.J.M. Van den and P. Nijkamp (1991), 'A general dynamic economic-ecological model for regional sustainable development', *Journal of Environmental Systems*, **20**, 89–114.

Bergh, J.C.J.M. Van den and P. Nijkamp (1997), 'Optimal growth, coordination, and sustainability in the spatial economy', mimeo, Free University, Amsterdam, the Netherlands.

Bergh, J.C.J.M. Van den and P. Nijkamp (1998), 'A multiregional perspective on growth and environment: the role of endogenous technology and trade', *Annals of Regional Science*, **32**, 115–31.

Bhattacharyya, A., T.R. Harris and R. Narayanan (1995), 'Allocative efficiency of rural Nevada water systems: a hedonic shadow cost function approach', *Journal of Regional Science*, **35**, 485–501.

Bottazzi, G., G. Dosi, G. Faglio and A. Secchi (2007), 'Modeling industrial evolution in geographical space', *Journal of Economic Geography*, **7**, 651–72.

Brundtland, G.H. (1987), *Our Common Future*, Oxford: Oxford University Press.

Burn, D.H. (1994), 'Hydrologic effects of climate change in west-central Canada', *Journal of Hydrology*, **160**, 53–70.

Calmette, M. and I. Pechoux (2006), 'Regional agglomeration of major risky activities and environmental policies', *Canadian Journal of Regional Science*, **29**, 177–94.
Calzadilla, A., F. Pauli and R. Roson (2007), 'Climate change and extreme events: an assessment of economic implications', *International Journal of Ecological Economics and Statistics*, **7**, 5–28.
Campbell, H.J. and I. Masser (1995), *GIS and Organisations: How Effective are GIS in Practice?*, London: Taylor & Francis.
Carlton, D.W. and G.C. Loury (1980), 'The limitations of Pigouvian taxes as a long-run remedy for externalities', *Quarterly Journal of Economics*, **94**, 559–66.
Carlton, D.W. and G.C. Loury (1986), 'The limitations of Pigouvian taxes as a long-run remedy for externalities: an extension of results', *Quarterly Journal of Economics*, **101**, 631–4.
Cohen, S. (2006), *Understanding Environmental Policy*, New York: Columbia University Press.
Cole, M.A. and R.J. Elliott (2007), 'Do environmental regulations cost jobs? An industry-level analysis of the UK', *BE Journal of Economic Analysis and Policy*, **7**, 1–25.
Duffy-Deno, K.T. (1992), 'Pollution abatement expenditures and regional manufacturing activity', *Journal of Regional Science*, **32**, 419–36.
Gabe, T.M. (2007), 'Local economic instability and business location: the case of Maine', *Land Economics*, **83**, 398–411.
Garofalo, G. and D.M. Malhotra (1995), 'Effect of environmental regulations on state-level manufacturing capital formation', *Journal of Regional Science*, **35**, 201–16.
Geoghegan, J., L.A. Wainger and N.E. Bockstael (1997), 'Spatial landscape indices in a hedonic framework: an ecological economics analysis using GIS', *Ecological Economics*, **23**, 251–64.
Ghosh, S., R. Vale and B. Vale (2006), 'Domestic energy sustainability of different residential patterns: a New Zealand approach', *International Journal of Sustainable Development*, **9**, 16–37.
Giaoutzi, M. and P. Nijkamp (1994), *Models for Regional Sustainable Development*, Avebury: Ashgate.
Girard, L.F. and P. De Toro (2007), 'Integrated spatial assessment: a multicriteria approach to sustainable development of cultural and environmental heritage in San Marco dei Cavoti, Italy', *Central European Journal of Operations Research*, **15**, 281–99.
Gleick, P.H. (1987), 'Regional hydrologic consequences of increases in atmospheric carbon dioxide and other trace gases', *Climatic Change*, **10**, 137–61.
Gress, D.R. and J.P.H. Poon (2007), 'Firm networks and Korean subsidiaries in the United States', *Growth and Change*, **38**, 396–418.
Guhathakurta, S. and P. Gober (2007), 'The impact of the Phoenix urban heat island on residential water use', *Journal of the American Planning Association*, **73**, 317–29.
Gutman, P. (2007), 'Ecosystem services: foundations for a new rural–urban compact', *Ecological Economics*, **62**, 383–87.
Hanak, E. and A. Chen (2007), 'Wet growth: effects of water policies on land use in the American West', *Journal of Regional Science*, **47**, 85–108.
Hosoe, M. and T. Naito (2006), 'Trans-boundary pollution transmission and regional agglomeration effects', *Papers in Regional Science*, **85**, 99–119.
Hu, J., C. Huang and W. Chu (2004), 'Bribery, hierarchical government, and incomplete environmental enforcement', *Environmental Economics and Policy Studies*, **6**, 177–96.
Hunter, L.M. (2006), 'Household strategies in the face of resource scarcity in coastal Ghana: are they associated with development priorities?', *Population Research and Policy Review*, **25**, 157–74.
Islam, S.M.N., M. Munasinghe and M. Clarke (2003), 'Making long-term economic growth more sustainable: evaluating the costs and benefits', *Ecological Economics*, **47**, 149–66.
Kahn, J.R. (1998), *The Economic Approach to Environmental and Natural Resources*, 2nd edn, Fort Worth, TX: Dryden Press.
Kim, T. and M.W. Horner (2003), 'Exploring spatial effects on urban housing duration', *Environment and Planning A*, **35**, 1415–29.
Kim, T., M.W. Horner and R.W. Marans (2005), 'Life cycle and environmental factors in selecting residential and job locations', *Housing Studies*, **20**, 457–73.
Kinnaman, T.C. (2006), 'Policy watch: examining the justification for residential recycling', *Journal of Economic Perspectives*, **20**, 219.
Krugman, P.R. (1991), 'Increasing returns and economic geography', *Journal of Political Economy*, **99**, 483–99.
Lambert, T. and C. Boerner (1997), 'Environmental inequity: economic causes, economic solutions', *Yale Journal on Regulation*, **14**, 195–234.
Lee, M. (2007), 'The effect of environmental regulations: a restricted cost function for Korean manufacturing industries', *Environment and Development Economics*, **12**, 91–104.
Lo, J.C., A.K.H. Lau and F. Chen (2007), 'Urban modification in a mesoscale model and the effects on the local circulation in the Pearl River Delta region', *Journal of Applied Meteorology and Climatology*, **46**, 457–76.

Longley, P.A. (1998), 'GIS and the development of digital urban infrastructure', *Environment and Planning B*, **25**, 53–6.
McDavid, J.C. (2000), 'Alternative service delivery in Canadian local governments: the costs of producing solid waste management services', *Canadian Journal of Regional Science*, **23**, 157–74.
Millimet, D.L. and J.A. List (2004), 'The case of the missing pollution haven hypothesis', *Journal of Regulatory Economics*, **26**, 239–62.
Ming, J., L. Xian-Guo, X. Lin-Shu, C. Li-Juan and T. Shouzheng (2007), 'Flood mitigation benefit of wetland soil: a case study in Momoge National Nature Reserve in China', *Ecological Economics*, **61**, 217–23.
Mitchell-Weaver, C. (1992), 'Public–private partnerships, innovation networks, and regional development in southwestern Pennsylvania', *Canadian Journal of Regional Science*, **15**, 273–88.
Morisugi, H. and E. Ohno (1995), 'Proposal of a benefit incidence matrix for urban development projects', *Regional Science and Urban Economics*, **25**, 461–81.
Nelson, G.C. (ed.) (2002), 'Spatial analysis for agricultural economists', Special Issue, *Agricultural Economics*, **27** (3).
Nelson, G., A. De Pinto, V. Harris and S. Stone (2004), 'Land use and road improvements: a spatial perspective', *International Regional Science Review*, **27**, 297–325.
Nijkamp, P., M. van der Burch and G. Vindigni (2002), 'A comparative institutional evaluation of public–private partnerships in Dutch urban land-use and revitalisation', *Urban Studies*, **39**, 1865–80.
Partridge, M.D. (2007), 'Rural economic development prospects in a high energy cost environment', *Journal of Regional Analysis and Policy*, **37**, 44–7.
Pfaff, A., J. Robalino, R. Walker, S. Aldrich, M. Caldas, E. Reis, S. Perz, C. Bohrer, E. Arima, W. Laurance and K. Kirby (2007), 'Road investments, spatial spillovers, and deforestation', *Journal of Regional Science*, **47**, 109–23.
Piore, M. and C. Sable (1984), *The Second Industrial Divide*, New York: Basic Books.
Polese, M. and R. Shearmur (2006), 'Why some regions will decline: a Canadian case study with thoughts on local development strategies', *Papers in Regional Science*, **85**, 23–46.
Portnov, B.A., M. Adhikari and M. Schwartz (2007), 'Urban growth in Nepal: does location matter?', *Urban Studies*, **44**, 915–37.
Press, K. (2007), 'When does defection pay? The stability of institutional arrangements in clusters', *Journal of Economic Interaction and Coordination*, **2**, 67–84.
Quaas, M.F. (2007), 'Pollution-reducing infrastructure and urban environmental policy', *Environment and Development Economics*, **12**, 213–34.
Ruth, M. (2006a), 'Introduction', in M. Ruth (ed.), *Smart Growth and Climate Change*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, pp. 3–8.
Ruth, M. (2006b), 'A summary of lessons and options', in M. Ruth (ed.), *Smart Growth and Climate Change*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, pp. 393–400.
Ruth, M., K. Donaghy and P. Kirshen (2006), 'Introduction', in M. Ruth, K. Donaghy and P. Kirshen (eds), *Regional Climate Change and Variability*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, pp. 1–29.
Scott, A. and M. Storper (eds) (1986), *Production, Work, Territory*, London: Allen & Unwin.
Shen, G. (2002), 'Measuring accessibility of housing to public-community facilities using geographical information systems', *Review of Urban and Regional Development Studies*, **14**, 235–55.
Sicotte, D. and S. Swanson (2007), 'Whose risk in Philadelphia? Proximity to unequally hazardous industrial facilities', *Social Science Quarterly*, **88**, 515–34.
Silva, C.L. and H.C. Jenkins-Smith (2007), 'The precautionary principle in context: US and EU scientists' prescriptions for policy in the face of uncertainty', *Social Science Quarterly*, **88**, 640–64.
Smith, C.L. (2007), 'Economic deprivation and environmental inequality in postindustrial Detroit: a comparison of landfill and Superfund site locations', *Organization and Environment*, **20**, 25–43.
Smith, J.B. and R.O. Mendelsohn (eds) (2006), *The Impact of Climate Change on Regional Systems*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.
Stevens, D.L. and A.R. Olsen (2004), 'Spatially balanced sampling of natural resources', *Journal of the American Statistical Association*, **99**, 262–78.
Stone, B. (2006), 'Physical planning and urban heat island formation: how cities change regional climates', in M. Ruth (ed.), *Smart Growth and Climate Change*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, pp. 318–41.
Tobin, J. (1969), 'A general equilibrium approach to monetary theory', *Journal of Money, Credit, and Banking*, **1**, 15–29.
Troesken, W. (2002), 'The limits of Jim Crow: race and the provision of water and sewerage services in American cities, 1880–1925', *Journal of Economic History*, **62**, 734–72.
Tsou, M. (2004), 'Integrating web-based GIS and image processing tools for environmental monitoring and natural resource management', *Journal of Geographical Systems*, **6**, 155–74.

Uimonen, S. (2001), 'The insufficiency of Pigouvian taxes in a spatial general equilibrium model', *Annals of Regional Science*, **35**, 283–98.
Velarde, S.J., Y. Malhi, D. Moran, J. Wright and S. Hussain (2005), 'Valuing the impacts of climate change on protected areas in Africa', *Ecological Economics*, **53**, 21–33.
Verhoef, E.T. (2003), 'Inside the queue: hypercongestion and road pricing in a continuous time-continuous place model of traffic congestion', *Journal of Urban Economics*, **54**, 531–65.
Verhoef, E.T. and P. Nijkamp (2000), 'Spatial dimensions of environmental policies for transboundary externalities: a spatial price equilibrium approach', *Environment and Planning A*, **32**, 2033–55.
Verhoef, E.T., J.C.J.M. van den Bergh and K.J. Button (1997), 'Transport, spatial economy, and the global environment', *Environment and Planning A*, **29**, 1195–213.
Walker, R. (2004), 'Theorizing land-cover and land-use changes: the case of tropical deforestation', *International Regional Science Review*, **27**, 247–70.
Walker, R. and T.E. Smith (1993), 'Tropical deforestation and forest management under the system of concession logging: a decision-theoretic analysis', *Journal of Regional Science*, **33**, 387–419.
Wallis, A., A. Richards, K. O'Toole and B. Mitchell (2007), 'Measuring regional sustainability: lessons to be learned', *International Journal of Environment and Sustainable Development*, **6**, 193–207.
Wellisch, D. and W.F. Richter (1995), 'Internalizing interregional externalities by regionalization', *Regional Science and Urban Economics*, **25**, 685–704.
Wen, Z., K. Zhang, B. Du, Y. Li and W. Li (2007), 'Case study on the use of genuine progress indicator to measure urban economic welfare in China', *Ecological Economics*, **63**, 463–75.
Wirl, F. (2006), 'Pollution thresholds under uncertainty', *Environment and Development Economics*, **11**, 493–506.
Ye, M. and A.M.J. Yezer (1997), 'Where will we put the garbage? Economic efficiency versus collective choice', *Regional Science and Urban Economics*, **27**, 47–66.

PART IV

REGIONAL GROWTH AND DEVELOPMENT MEASUREMENT METHODS

16 Measuring agglomeration

Ryohei Nakamura and Catherine J. Morrison Paul

16.1 Introduction

The existence of agglomeration economies is crucial both for explaining the size and distribution of modern cities and for understanding their growth and development. Agglomeration economies are also important policy issues for regional municipalities and national governments, because they engender industrial clustering. Academic studies of regional clustering typically focus on measuring the extent of agglomeration and its associated economies, or determining the mechanisms underlying, and the effects of, agglomeration economies. In this chapter we will focus on the former, with some overview of the latter.

Before discussing the measurement of agglomeration, it is necessary to define economic agglomeration. The term 'agglomeration' is often used interchangeably with 'specialization' or 'concentration'. Brülhart (1998, p. 776), however, suggests that specialization and agglomeration involve immobile and mobile factors, respectively.[1] That is, specialization refers to industrial composition in a specific region in which some industries are agglomerated compared to their national counterparts, which is a relative rather than an absolute measurement of agglomeration. In turn, agglomeration typically refers to spatial concentration of economic activity in a limited area, while (spatial) concentration often applies to the spatial distribution of specific industries. For example, the food industry is relatively evenly distributed among regions while the textile industry tends to be concentrated in particular regions.[2] In this chapter we will consider agglomeration to be inclusive of both specialization and concentration.

We will focus on the measurement of economic agglomeration in the context of the clustering of regional economic activity. We first discuss various agglomeration measures that have been proposed in the literature, to provide alternative methodologies for the direct measurement of agglomeration. We then briefly review the estimation of the determinants or sources of agglomeration, and the resulting agglomeration economies or productivity effects of agglomeration, both of which involve methods of indirect measurement. In conclusion we highlight some topics that will be important to address in future studies of agglomeration economies.

16.2 Direct measurement of agglomeration

Various measures may be constructed to represent clustering or agglomeration. Measures may, for example, be aggregated across regions or industries, and be computed relative to broader aggregations in terms of ratios or differences. They may also be expressed in terms of the concentration of employment, plants or output, and may represent agglomeration in terms of own-industry or between-industry effects. In the following sections we identify and interpret these different dimensions along which agglomeration measures may be computed.

Industrial localization in terms of employment

We first consider measures of the spatial distribution or industrial localization of industry i in terms of employment, which reflect the geographic concentration of industry i employment across regions. Assuming that there are J regional (or geographical) units and I industries in a country, with the number of employees of industry i in region j denoted by x_{ij}, constructing such a measure involves characterizing employment in industry i region j as a share of employment in industry i over all regions.

Representing the spatial distribution of industry i first requires measuring spatial concentration (denoted C) for each region j based on employment density:

$$s_{i1}^C, s_{i2}^C, \ldots, s_{ij}^C, \ldots, s_{iJ}^C, \text{ where}$$

$$s_{ij}^C = \frac{x_{ij}}{\sum_{j=1}^{n} x_{ij}} = \frac{x_{ij}}{x_{i*}},\ i = 1,\ldots, I;\ j = 1, \ldots, J. \tag{16.1}$$

s_{ij}^C therefore reflects the employment share of industry i, region j, in total (national) industry i employment, or the concentration of industry i in region j relative to all regions. The spatial distribution of employment by region for all industries is then represented by aggregating the s_{ij}^C measures across all industries, resulting in:

$$s_{*1}, s_{*2}, \cdots, s_{*j}, \ldots, s_{*n}, \text{ where}$$

$$s_{*j} = \frac{\sum_{i=1}^{I} x_{ij}}{\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}} = \frac{x_{*j}}{x_{**}} \tag{16.2}$$

shows the relative size of economic activity in terms of each region's total employment share.

One summary measure of geographic concentration, based on the employment share measures in equation (16.1), is computed as the sum of squares of s_{ij}^C over all regions:

$$H_i^C = \sum_{j=1}^{J} (s_{ij}^C)^2. \tag{16.3}$$

This is a form of Hirschman–Herfindahl index, which is equal to one if the industry is fully concentrated in one region and approaches zero if the industry is evenly distributed with very small shares over a great number of regions.[3]

Another measure of agglomeration combines the spatial concentration measure for industry i in equation (16.1) with that for all industries in equation (16.2) to compute the concentration of industry i in region j relative to the concentration of all industries (or economic size) in region j, compared to the nation as a whole:

$$LQ_{ij}^C = \frac{s_{ij}^C}{s_{*j}} = \frac{x_{ij}/x_{i*}}{x_{*j}/x_{**}},\ j = 1,\ldots,J. \tag{16.4}$$

That is, this form of location quotient (LQ) reflects the percentage (share) of industry i's productive activity in region j relative to the percentage (share) of total productive activity in region j, expressed in terms of employment. Computing the mean of these LQ_{ij}^C measures:

$$LOC_i^C = \frac{1}{J}\sum_{j=1}^{J}\frac{s_{ij}^C}{s_{*j}}, \tag{16.5}$$

results in an average concentration index showing the degree of localization (*LOC*) of industry *i*, often called the 'industrial localization rate' in the field of economic geography.

As an alternative to measuring relative concentration of industry *i* compared to all industries as a ratio, one can compute as the difference between equations (16.1) and (16.2): $s_{ij}^C - s_{*j}$. If this difference is positive for region *j*, region *j* is more specialized in industry *i* compared to overall industries, or the employment share of industry *i* in region *j* is high relative to the share of total employment in region *j*, analogously to LQ_{ij}^C>1.

Summing the absolute value or the squared sum of this measure over regions expresses the degree of location deviation or spatial concentration of industry *i* as the '*G*' measure of Audretsch and Feldman (1996):

$$G_i^C = \frac{1}{J}\sum_{j=1}^{J}\left|s_{ij}^C - s_{*j}\right| \tag{16.6}$$

or

$$G_i^{C_2} = \frac{1}{J}\sum_{j=1}^{J}(s_{ij}^C - s_{*j})^2. \tag{16.7}$$

These measures, sometimes called dissimilarity measures,[4] take a value near zero if the spatial distribution of industry *i* is similar to that of all industries.

Another way the information in the measures s_{ij}^C and s_{*j} can be summarized is in terms of a location Gini coefficient that represents the degree of geographical concentration – a spatial application of the Gini coefficient often used in income distribution studies to measure inequality. This coefficient is defined as the area between a 45-degree line for equal distribution and a Lorenz concentration curve for the observed distribution, so its value measures relative concentration with zero indicating an equal distribution.

More specifically, the Gini location coefficient is calculated by first ranking the concentration index (16.4) in ascending order, cumulating the numerator and the denominator, and then plotting these values on the vertical and horizontal axes, respectively, to derive a Lorenz curve. The Gini coefficient is then computed as twice the area between a 45-degree line and this Lorenz curve.[5]

This computation can be formalized, as in Aiginger et al. (1999) as:

$$GINI_i^C = \frac{0.5 - \frac{1}{2J}\sum_{j=1}^{J}(s_{ij-1}^C + s_{ij}^C)}{0.5\left(1 - \frac{1}{J}\right)} \tag{16.8}$$

or

$$GINI_i^C = \frac{2}{J^2\left(\frac{1}{J}\sum_{j=1}^{J}\frac{s_{ij}^C}{s_{*j}}\right)}\sum_{j=1}^{J}\lambda_i\left|\frac{s_{ij}^C}{s_{*j}} - \frac{1}{J}\sum_{j=1}^{J}\frac{s_{ij}^C}{s_{*j}}\right|, \tag{16.8'}$$

so $GINI_i^C$ measures the degree to which the percentage distribution of industry *i*'s employment across regions coincides with the percentage distribution of national employment across regions. This coefficient takes values between zero and one, where zero means employment in industry *i* is distributed over regions identically to that of total employment.

Although the Gini location coefficient provides an aggregate measure of the concentration of industry employment distribution across regions, its interpretability is limited because the same coefficient values can stem from different spatial employment distributions. That is, it explains the degree of concentration for an industry but does not show how firms or plants are distributed among regions.

Regional specialization in terms of employment

Another perspective on agglomeration is in terms of regional specialization, defined as the share of industry *i*'s employment relative to total industry employment in a specific region *j*, by contrast to the share of region *j*'s employment relative to total (national) employment in industry *i* as in the previous section. That is, the level of specialization (denoted *S*) in region *j* with respect to industry *i* is given by:

$$s_{ij}^S = \frac{x_{ij}}{\sum_{i=1}^{I} x_{ij}} = \frac{x_{ij}}{x_{*j}}, i = 1,\ldots, I;\ j = 1,\ldots, J, \tag{16.9}$$

where the denominator is aggregated over industries rather than over regions as for industrial localization. Similarly to the industrial concentration measures for spatial distribution discussed above, the regional specialization measures reflecting the industrial composition or structure for each region *j* (based on employment levels) are therefore:

$$s_{1j}^S,\ s_{2j}^S, \cdots, s_{ij}^S, \ldots, s_{Ij}^S,$$

and the Hirschman–Herfindahl index for regional specialization can be calculated as:

$$H_j^S = \sum_{i=1}^{I} (s_{ij}^S)^2. \tag{16.10}$$

Further, at a national level industrial composition is represented by

$$s_{1*}, s_{2*}, \cdots, s_{i*}, \ldots, s_{I*}, \text{ where}$$

$$s_{i*} = \frac{\sum_{j=1}^{J} x_{ij}}{\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}} = \frac{x_{i*}}{x_{**}}, \tag{16.11}$$

so the regional specialization index relative to national industrial composition is expressed as:

$$LQ_{ij}^S = \frac{s_{ij}^S}{s_{i*}} = \frac{x_{ij}/x_{*j}}{x_{i*}/x_{**}},\ i = 1,\ldots,I. \tag{16.12}$$

That is, this form of location quotient represents the specialization of industry *i* in region *j* relative to the specialization of industry *i* in all regions. The average of these location quotients across industries can in turn be computed as:

$$LOC_j^S = \frac{1}{I}\sum_{i=1}^{I}\frac{s_{ij}^S}{s_{i*}}, \tag{16.13}$$

where $LOC_j^S > 1$ indicates a high relative level of regional specialization for region *j*.

One can alternatively sum the absolute differences of regional specialization and national composition instead of the ratios. This results in a specialization index analogous to the *G* index in equation (16.6) for spatial concentration:[6]

$$G_j^S = \frac{1}{I}\sum_{i=1}^{I}\left|s_{ij}^S - s_{i*}\right|, \tag{16.14}$$

which takes a value of zero rather than one if region *j* has an industrial composition identical to the nation and takes a maximum value of 2/I if the region has no industries in common with the nation. Since this index shows the deviation of industrial structure from the national average, it is sometimes called a dissimilarity index of regional specialization.[7]

Krugman's (1991) measure of regional specialization represents relative industrial composition by bilateral comparison of two regions, rather than of a given region relative to the nation. This measure:

$$K_{jk}^S = \sum_{i=1}^{I}\left|s_{ij}^S - s_{ik}^S\right|, \tag{16.15}$$

takes a value of zero if the two regions have identical industrial composition. Many researchers have used this index, or an adaptation of this measure, to compare specialization between regions or nations.[8]

It is also possible to compute a Gini coefficient for regional specialization from the information contained in the s_{ij}^S measures, similarly to equations (16.8) or (16.8)'. In this case, aggregating across industries instead of regions to reflect the deviation from equal distribution of industries in a particular region results in:

$$GINI_j^S = \frac{0.5 - \frac{1}{2I}\sum_{i=1}^{I}(s_{i-1j}^S + s_{ij}^S)}{0.5\left(1 - \frac{1}{I}\right)} \tag{16.16}$$

or

$$GINI_j^S = \frac{2}{I^2\left(\frac{1}{I}\sum_{i=1}^{I}\frac{s_{ij}^S}{s_{i*}}\right)}\sum_{i=1}^{I}\lambda_i\left|\frac{s_{ij}^S}{s_{i*}} - \frac{1}{I}\sum_{i=1}^{I}\frac{s_{ij}^S}{s_{i*}}\right|, \tag{16.16'}$$

where λ_i indicates the position of industry *i* in the ranking, in descending order, of location quotient LQ_{ij}^S. This index shows the inequality of the distribution of industrial

composition in region *j* compared to the national distribution. If the industrial composition of region *j* coincides with the national composition the Gini coefficient takes a value of zero. A higher value reflects greater specialization of region *j* relative to the nation.

Location indexes in terms of plant size

In the two previous subsections concentration or specialization were expressed in terms of employment. However, agglomeration may also be measured in terms of the number of firms or plants locating in a particular area, which may provide a different perspective on the extent and causes of agglomeration.

Consider, for example, a region with 100 small plants in industry *i* with 50 employees each, compared to a region with two large plants in industry *i* with 2500 employees each. In the former case one might expect that the clustering of 100 plants in the same industry is the result of agglomeration economies, perhaps due to natural advantages or mutual externalities (as discussed further below). The latter case likely results from large plant-level scale economies in industry *i*, which leads to very few plants in the region even with the same amount of employment.

In both these cases the industry *i* employment-based location quotients for this region would have the same value. Further, the employment-based Gini coefficient would have a high value indicating spatial concentration of industry *i*. These measures therefore do not distinguish concentration of firms from that of employment.

A simple way to disentangle spatial concentration of plants and industrial specialization of employment is to calculate and compare the plant-based location quotients or indexes to the employment-based indexes. That is, if y_{ij} denotes the number of industry *i* plants in region *j*, the location quotient representing regional specialization with respect to plants, $LQ_{ij}^{S(P)}$, is expressed as:

$$LQ_{ij}^{S(P)} = \frac{y_{ij}/y_{*j}}{y_{i*}/y_{**}}, \tag{16.17}$$

by contrast to the location quotient with respect to employment:

$$LQ_{ij}^{S(E)} = \frac{x_{ij}/x_{*j}}{x_{i*}/x_{**}}. \tag{16.18}$$

Therefore, if $LQ_{ij}^{S(P)} > LQ_{ij}^{S(E)} > 1$, region *j* is specialized with a concentration of relatively small plants in industry *i*, whereas $LQ_{ij}^{S(E)} > LQ_{ij}^{S(P)} > 1$ indicates that a few large-sized plants are located in the region.

To facilitate interpretation this comparison can be expressed in the ratio form:

$$\frac{LQ_{ij}^{S(E)}}{LQ_{ij}^{S(P)}} = \frac{x_{ij}/x_{*j}}{x_{i*}/x_{**}} \Big/ \frac{y_{ij}/y_{*j}}{y_{i*}/y_{**}} = \frac{x_{ij}/y_{ij}}{x_{*j}/y_{*j}} \Big/ \frac{x_{i*}/y_{i*}}{x_{**}/y_{**}}, \tag{16.19}$$

where the numerator is:

$$\frac{x_{ij}/y_{ij}}{x_{*j}/y_{*j}} = \frac{\textit{average plant size of industry } i \textit{ in region } j}{\textit{average plant size of all industries in region } j}, \tag{16.20}$$

with size defined in terms of employment, and the denominator is:

$$\frac{x_{i*}/y_{i*}}{x_{**}/y_{**}} = \frac{\textit{average plant size of industry } i \textit{ nationwide}}{\textit{average plant size of all industries nationwide}}. \tag{16.21}$$

Therefore, if $LQ_{ij}^{S(E)} > LQ_{ij}^{S(P)}$, so this ratio is greater than one, region j contains relatively large plants or has a regional monopolistic/oligopolistic environment. However, if $LQ_{ij}^{S(P)} > LQ_{ij}^{S(E)}$, so the ratio falls short of one, region j contains relatively small plants or has a regional competitive environment.

One might also compute a Gini coefficient to represent deviations from equal distribution for the agglomeration problem expressed in terms of plants. However, Alecke et al. (2006) and others have shown that such a measure has interpretation problems because it measures the concentration of economic activity from both internal scale economies (that is, 'concentration' within a firm), and external scale economies or natural advantages (that is, concentration from firm clustering). To distinguish between these two causes of concentration one must therefore use an alternative measure – an agglomeration index accommodating plant size distribution – as described below in the section on the Ellison and Glaeser index. Before moving to such a measure, however, it is useful to consider one more adaptation of location coefficients or indexes to be expressed in terms of output levels.

Location indexes in terms of output levels

Location quotients are sometimes used as a regional specialization index to identify export industries in analyses of international trade. Such a measure, called a Balassa index, indicates regional comparative advantages among nations by a country's share of industry i world exports relative to its share of total world exports.[9] In this context, analogously to the location quotients with respect to plant numbers and employment in equations (16.17) and (16.18), the regional specialization index ρ_{ij} is defined in terms of industry i's regional output q_{ij}:

$$\rho_{ij} = \frac{q_{ij}/q_{*j}}{q_{i*}/q_{**}}. \tag{16.22}$$

Industry i in region j is thus regarded as an export sector if the industry i share of output in region j is greater than the national average. That is, region j is relatively specialized in production of industry i output and $\rho_{ij} > 1$.

However, it is possible for this measure to be biased if the ρ_{ij} index is directly applied to identify the export sector in a region. For example, say industry i is specialized in the nation as a whole, and the output of industry i is exported to foreign countries. ρ_{ij} will then be underestimated because the national economy is an open economic system with imports and exports. That is, the nation itself may be specialized globally as an exporter of the output of industry i rather than being a closed system. In reverse, if industry i is an import industry within the nation ρ_{ij} will be overestimated.

More specifically, imagine a basic input–output table. A balance equation characterizing total versus domestic demand can be written as $q_{i*} = q_{i*}^{D} + q_{i*}^{X} - q_{i*}^{M}$, where q_{i*}^{D} is domestic intermediate demand, q_{i*}^{X} is exports to foreign countries, and q_{i*}^{M} is imports from foreign countries. The specialization index of industry i at the national level is then defined, taking the export and import balance into account, as:

$$\rho_{i*} = \frac{q_{i*}/q_{**}}{q_{i*}^{D}/q_{**}^{D}} = \frac{q_{i*}/q_{**}}{(q_{i*} - q_{i*}^{X} + q_{i*}^{M})/(q_{**} - q_{**}^{X} + q_{**}^{M})}. \tag{16.23}$$

This national specialization measure can be used to adjust the location quotient in (16.22) to eliminate the bias associated with trade patterns, resulting in the modified location coefficient:

$$\hat{\rho}_{i*} = \frac{q_{ij}/q_{*j}}{q^D_{i*}/q^D_{**}} = \rho_{i*}\frac{q_{ij}/q_{*j}}{q_{i*}/q_{**}} = \rho_{i*}\,\rho_{ij}. \tag{16.24}$$

The domestic demand for region j, q^D_{ij}, can in turn be calculated as:

$$q^D_{ij} = \frac{q_{ij}}{\hat{\rho}_{ij}} = \frac{q_{ij}}{\rho_{i*}\rho_{ij}} = q_{*j}\left(\frac{q^D_{i*}}{q^D_{**}}\right), \tag{16.25}$$

where net exports (exports minus imports) are $q_{ij} - q^D_{ij}$.[10]

The Ellison and Glaeser index

Standard agglomeration measures can also be modified to accommodate plant size distribution. That is, as discussed above, if a small number of large firms locate in a few regions due to plant-level scale economies employment will be concentrated in those regions even though there are few firms (or plants), which should be recognized in the agglomeration measure to facilitate its interpretation.

Recall that the value of the index specified in equation (16.7) to show industry i's geographic concentration,

$$G^2_i = \frac{1}{J}\sum_{j=1}^{J}(s^C_{ij} - s_{*j})^2,$$

is zero if employment in industry i and total employment have an identical geographic distribution and increases the more the respective industry is concentrated in a few regions. In the case of a few large localized firms, G^2_i will take a high value in terms of industry i employment in the absence of agglomeration in terms of firms' spatial concentration. To identify agglomeration economies implied by such industrial distribution, plant size must be incorporated into this measure.[11]

Ellison and Glaeser (1994, 1997) recognize the dependency between industrial distribution and geographic concentration by developing a probabilistic location model based on 'throwing darts' at plants in a country map. If there are no natural advantages or spillovers between firms, the probability of locating in region j depends solely on the geographical size of the region. However, in the presence of such spatial externalities agglomeration should be captured by the agglomeration measure.

Ellison and Glaeser (1994) first define a normalized G measure controlling for the distribution of national employment in industry i, denoted raw geographic concentration:

$$G^{(EG)}_i = \frac{\sum_{j=1}^{J}(s^C_{ij} - s_{*j})^2}{1 - \sum_{j=1}^{J}(s_{*j})^2}, \tag{16.26}$$

where the denominator takes a value of one if total employment in the industry is evenly distributed across regions. In turn, plant size distribution in industry i is measured using a Hirschman–Herfindahl index based on the number of plants, rather than regions or industries as in equations (16.2) and (16.10):

$$H_i^P = \sum_{k=1}^{K} (z_{k\epsilon i})^2, \tag{16.27}$$

where $z_{k\epsilon i}$ denotes the employment share of plant k in industry i and K is the number of plants in industry i. If all plants are the same size with respect to employment, the inverse of H_i^P collapses to the number of plants, K. The more uneven the plant size distribution, or the higher the level of industrial concentration, the smaller is H_i^P.

Ellison and Glaeser (1994) use the expected value of the raw concentration measure, $E[G_i^{(EG)}]$, given by:

$$E[G_i^{(EG)}] = \gamma_i(1 - H_i^P) + H_i^P,$$

to derive the estimator γ_i representing the excess of raw geographic concentration relative to productive concentration with respect to industry i:

$$\hat{\gamma}_i^{(EG)} = \frac{G_i^{(EG)} - H_i^P}{1 - H_i^P} = \frac{\sum_{j=1}^{J} (s_{ij}^C - s_{*j})^2 - \left(1 - \sum_{j=1}^{J} (s_{*j})^2\right) \sum_{k=1}^{K} (z_{k\epsilon i})^2}{\left(1 - \sum_{j=1}^{J} (s_{*j})^2\right)\left(1 - \sum_{k=1}^{K} (z_{k\epsilon i})^2\right)}, \tag{16.28}$$

where the numerator shows the difference between the degree of geographic concentration of industry i (the spatial Hirschman–Herfindahl index) and its expected value.[12] This index, typically called the Ellison and Glaeser index, indicates whether concentration is greater than the expected value of random location of firms (without suggesting a reason for the agglomeration). $\hat{\gamma}_i^{(EG)}$ is therefore interpreted as a combined measure of the strength of all agglomeration drivers such as natural advantages and spillovers among plants.[13]

Maurel and Sédillot (1999) propose a similar agglomeration index that focuses on spillovers from the proximity of same-industry plants. They define a binominal variable as $u_{kj} = 1$ if plant k locates in region j, and $u_{kj} = 0$ otherwise, which is a non-independent Bernouilli variable given that the probability of plant k locating in region j is $Pr(u_{kj} = 1) = s_{*j}$. That is, random locating behavior of plants among regions results in a pattern of employment shares that matches the aggregate, s_{*j}.

Maurel and Sédillot (1999) accordingly redefine the Ellison and Glaeser estimator as the interaction (correlation) between the locations of any pair of plants in industry i:

$$\gamma_i = Corr(u_{kj}, u_{lj}) \text{ for } k \neq j,^{14} \tag{16.29}$$

where $-1 < \gamma_i < 1$ describes the strength of spillovers within the industry. The probability that two plants in the same industry i arbitrarily locate in the same region is therefore:

$$p_i = \sum_{j=1}^{n} p_{ij} = \gamma_i\left(1 - \sum_{j=1}^{J} (s_{*j})^2\right) + \sum_{j=1}^{J} (s_{*j})^2, \tag{16.30}$$

where $p_{ij} = E[u_{kj}, u_{lj}] = Cov(u_{kj}, u_{lj}) + E[u_{kj}]E[u_{lj}] = \gamma_i s_{*j}(1 - s_{*j}) + (s_{*j})^2$.

They then develop an estimator of p_{ij}:

$$\hat{p}_{ij} = \frac{\sum_{k \neq l}^{K} z_{k \epsilon j} z_{l \epsilon j}}{\sum_{k \neq l}^{K} z_k z_l}, \tag{16.31}$$

where $k,l \in j$ means that plants k and l locate in region j. After some manipulations, this estimator is expressed as:

$$\hat{p}_i = \frac{\sum_{j=1}^{J} (s_{ij}^C)^2 - H_i^P}{1 - H_i^P}. \tag{16.32}$$

Substituting equation (16.32) instead of p_i into equation (16.30) results in the estimator:

$$\hat{\gamma}_i^{(MS)} = \frac{G_i^{(MS)} - H_i^P}{1 - H_i^P} = \frac{\sum_{j=1}^{J} (s_{ij}^C)^2 - \sum_{j=1}^{J} (s_{*j})^2 - \left(1 - \sum_{j=1}^{J} (s_{*j})^2\right) \sum_{k=1}^{K} (z_{k \epsilon i})^2}{\left(1 - \sum_{j=1}^{J} (s_{*j})^2\right)\left(1 - \sum_{k=1}^{K} (z_{k \epsilon i})^2\right)}. \tag{16.33}$$

This agglomeration measure is preferable to $\hat{\gamma}_i^{(EG)}$ because it is directly derived from a 'dartboard' model, although both $\hat{\gamma}_i^{(EG)}$ and $\hat{\gamma}_i^{(MS)}$ are unbiased estimators of γ_i.

A modification for industry co-agglomeration

All of the agglomeration measures we have discussed so far involve the location of firms within the same industry. Ellison and Glaeser also consider inter-industry agglomeration, denoted co-agglomeration. Co-agglomeration exists if externalities induce firms in two industries to locate closer because they are vertically related with respect to intermediate input transactions. For example, say an industry in a two-digit manufacturing classification has R sub-industries, such as a four-digit classification. One might expect externalities to exist between establishments or plants in the different four-digit sub-industries within the same two-digit industry, as well as in the same four-digit sub-industry as measured by $\hat{\gamma}_{i \epsilon r}^{(EG)}$.

To capture agglomeration from such externalities, Ellison and Glaeser propose a co-agglomeration measure representing the difference between the degree of spatial concentration for a four-digit group of industries and the average weighted degree of spatial concentration for each industry group. This measure is written as:

$$\hat{\gamma}_{i \epsilon r}^{(EG)} = \frac{G_{i \epsilon r}^{(EG)} - \sum_{i \epsilon r}^{R} w_{i \epsilon r}^2 H_{i \epsilon r}^P - \sum_{i \epsilon r}^{R} \hat{\gamma}_{i \epsilon r} w_{i \epsilon r}^2 (1 - H_{i \epsilon r}^P)}{1 - \sum_{i \epsilon r}^{R} w_{i \epsilon r}^2}, \tag{16.34}$$

where $H_{i \epsilon r}^P$ denotes the plant-level Hirchman–Herfindahl index of the rth sub-industry group, and $w_{i \epsilon r}$ is the share of the rth sub-industry group in the two-digit industry employment. $\hat{\gamma}_{i \epsilon r}^{(EG)}$ thus reflects the extent to which the locations of firms in a particular industry are correlated.[15]

This concentration index is sophisticated, but requires data for the distribution of employment at the plant level that often are unavailable. In Ellison and Glaeser (1994, 1997), for example, employment data was only available by region so the Hirschman–Herfindahl index was taken from a separate data source.[16] However, if data limitations can be overcome, plant-level or geographically detailed data may be exploited to construct such measures. Empirical studies that have adopted or extended the Ellison and Glaeser framework to examine the geographic concentration of economic activity include Devereux et al. (2004) for the UK, Maurel and Sédillot (1999) for France, Barrios et al. (2003) for small European countries (Belgium, Ireland and Portugal), Braunerhjelm and Borgman (2004) for Sweden, Bertinelli and Decrop (2005) for Belgian manufacturing industries, Alecke et al. (2006) for German manufacturing plants, Lafourcade and Mion (2007) for Italy and Tokunaga and Akune (2005) for Japanese manufacturing industries.

The spatial scale of agglomeration

Bertinelli and Decrop (2005) evaluate measures of both the Ellison and Glaeser index and the location Gini coefficient, based on five conditions identified by Duranton and Overman (2002). These conditions are that an index of spatial concentration should: (1) be comparable across industries; (2) control for overall agglomeration of manufacturing; (3) distinguish spatial from industrial concentration; (4) be unbiased with respect to the degree of spatial and industrial aggregation; and (5) indicate statistical significance. The Gini coefficient fulfils the first two criteria, as well as being readily calculated based on generally available data. The Ellison and Glaeser index, although in some ways a significant improvement over earlier agglomeration or concentration measures, also only satisfies the first two criteria. In particular, it is limited because it does not recognize the spatial range of natural advantage or spillover effects, which will affect the density of agglomeration and the sensitivity of the localization measure to the definition of geographic area.

More specifically, the Ellison Glaeser index does not treat distance decay effects across bordering regions and depends on the spatial aggregation unit, so once a geographical unit is chosen and the concentration index is calculated it is difficult to compare the results with those based on other geographic scales. Two approaches have been developed to address such issues. One approach uses non-parametric methods to construct a continuous distance-based concentration index based on estimates of a Kernel density function. The second recognizes spatial autocorrelation.

Duranton and Overman (2002, 2005) and Marcon and Puech (2003) independently proposed distance-based methods to accommodate spatial borders. Duranton and Overman (2002, 2005) treat space continuously and calculate the distribution of all pairwise distances between plants in the same industry. In contrast to area-based indexes, their distance-based index is unbiased with respect to geographical size and aggregation, facilitating statistical inference of deviations from random location.

Macron and Puech use a homogenous spatial Poisson process to examine industrial concentration in France. Their distance-based methodology is based on Besag's L function, derived from Ripley's K function that defines the spatial distribution of a point process via the number of neighbors divided by the average density (Besag, 1977; Ripley, 1976, 1977). Besag's (normalized) L function measures concentration by computing the average number of a firm's neighbors for each radius of a circle, and checking whether the

actual distribution of firms differs significantly from a random pattern (Marcon and Puech, 2003, p. 413).[17] Marcon and Puech's results support previous findings that show aggregate activity in France is concentrated around Paris.

Lafourcade and Mion (2007) instead construct modified indexes of spatial autocorrelation to circumvent the choice of spatial scale. They investigate whether the geographic distribution of manufacturing activities is related to plant size, and whether physical distance helps to explain location patterns, using the spatial autocorrelation index 'Moran's *I*' – one of the classical indexes of spatial autocorrelation or interdependencies developed by Moran (1950). This statistic compares the value of a continuous variable at a given location with that of the same variable at contiguous locations:

$$Moran's\ I = \frac{J\sum_{j=1}^{J}\sum_{k=1}^{K} w_{i,jk}(x_{ij}-\bar{x}_i)(x_{ik}-\bar{x}_i)}{\left(\sum_{j=1}^{J}\sum_{k=1}^{K} w_{i,jk}\right)\sum_{k=1}^{K}(x_{ik}-\bar{x}_i)^2}, \tag{16.35}$$

where

$$\bar{x}_i = \frac{1}{J}\sum_{j=1}^{J} x_{ij}$$

and $w_{i,jk}$ is the weight in the contiguity matrix (in its simplest form equal to one for all contiguous regions and zero otherwise). Moran's I is thus interpreted as the correlation coefficient between the $x_i s$ in region j and its surrounding regions.

Lafourcade and Mion (2007) further recognize that agglomeration measures should recognize distance between establishments and be independent of region size, and so adapt their spatial autocorrelation indicator to account for distance-based co-location patterns across geographical units. Using Italian census data on manufacturing industries, they find strong evidence of a positive relationship between plant size and concentration, such that large plants are more concentrated but less agglomerated in terms of spatial autocorrelation whereas small plants are more agglomerated and less concentrated.[18]

Regional diversity and agglomeration

Regional diversity is another important agglomeration indicator associated with the variety of economic activity. That is, 'urbanization economies' stimulate density or agglomeration due to aspects of urban diversity such as consumption, labor market and industrial diversity. Consumption diversity raises consumer's utility in urban areas by offering opportunities to choose a variety of differentiated goods and services. Labor market diversity implies offerings and demand for a variety of jobs and skills, often summarized by the term 'labor market pooling'. Industrial diversity may also increase the potential for both own- and cross-industry externalities.[19]

The inverse of the Hirschman–Herfindahl index defined in terms of regional specialization has been used as a measure of regional diversity, for example by Duranton and Puga (2000):

$$DIV_j^A = 1/\sum_{i=1}^{I}(s_{ij}^S)^2,$$

where DIV_j^A takes a value of I (the number of industries in the industrial classification) if industrial employment in region j is evenly distributed among all industries (there is maximum diversification). Henderson et al. (1995) propose a somewhat different diversity index that reflects sectoral diversity faced by industry i excluding own-industry effects:

$$DIV_{ij}^B = \sum_{i'=1,i'\neq i}^{I} (s_{i'j}^S)^2. \tag{16.36}$$

Combes (2000) further extends Henderson's definition to be the inverse of a Herfindahl index of sectoral concentration based on the share of all industries except one:

$$DIV_{ij}^C = \frac{1/\sum_{i'=1,i'\neq i}^{I} (x_{i'j}/x_{-i,j})^2}{1/\sum_{i'=1,i'\neq i}^{I} (x_{i'*}/x_{-i,*})^2}, \tag{16.37}$$

where the numerator is maximized when all industries except the own industry have the same size in region j.

Such Hirschman–Hirfindahl indexes may be regarded as modified versions of the Gibbs–Martin index (GMI) of diversification developed by Gibbs and Martin (1962) in the field of quantitative geography (although most recent papers in the agglomeration literature do not recognize this earlier literature):

$$GMI_j = 1 - \sum_{i=1}^{I} (x_{ij})^2 \Big/ \left(\sum_{i=1}^{I} x_{ij} \right)^2. \tag{16.38}$$

If the labor force in a region is concentrated wholly in one industry $GMI_j = 0$, and if it is uniformly distributed throughout the whole industry (maximum diversification) GMI_j approaches one. If the GMI_j measure is based on the share s_{ij}^S instead of x_{ij} it takes a value of one minus the Hirschman–Herfindahl index because

$$\sum_{i=1}^{I} s_{ij} = 1.$$

Another index used to measure diversity is an entropy measure based on the second law of thermodynamics (the entropy law), which is also conceptually similar to the Hirschman–Herfindahl index. This index:

$$(Entropy\ Index)_j = -\sum_{i=1}^{I} \frac{x_{ij}}{x_{*j}} \log \frac{x_{ij}}{x_{i*}}, \tag{16.39}$$

is equal to zero if a region is completely specialized in one industry, and reaches its maximum value if employment is uniformly distributed among industries in region j.[20]

Many empirical studies have used Hirschman–Herfindahl or entropy indexes as diversity measures.[21] Diversity is sometimes regarded in these studies as the flipside of specialization, but this is not strictly true. As noted by Malliza and Ke (1993), diversity does not simply mean either the absence of specialization or a uniform distribution of activity among urban areas or metropolitan regions, but reflects the presence of multiple specializations (specialized diversity). In particular, in physical chemistry perfect diversity means the maximum entropy state based on the irreversible phenomenon known as the second

law of thermodynamics. However, in urban or regional economics equal shares of economic activity do not necessarily imply 'perfect' diversification.

16.3 Indirect measurement of agglomeration: sources and effects

In the previous section we described various indexes for measuring industrial specialization and regional concentration, which directly capture the degree of economic agglomeration. However, in some contexts the goal may instead be either to measure the sources (externalities or spillovers) or effects (performance or productivity) of the agglomeration behavior. Measures of these agglomeration sources or effects may be thought of as indirect agglomeration measures that can be used to evaluate the mechanisms underlying and resulting from agglomeration economies, respectively.[22]

Sources of agglomeration

As discussed already in Chapter 6, in his classic text Marshall (1920) identifies three sources of agglomeration economies: input sharing, labor market pooling and knowledge spillovers.[23] Input sharing involves scale economies in input production that enable downstream companies to purchase relatively inexpensive intermediate inputs from nearby companies.[24] Labor market pooling involves a concentration of workers that facilitates mutual learning and reduces the risks and costs of searching for workers or jobs. Knowledge spillovers involve interactions among people working in close proximity (even when improved information technology facilitates communication at greater distances) and turnover of skilled workers.

Two perspectives to knowledge spillovers have emerged in the literature, related to the standard division of agglomeration economies into localization economies, which involve the geographic concentration of same-industry firms, and urbanization economies, which involve density of diverse economic activity. Glaeser et al. (1992) suggest that increased concentration of a particular industry in a limited area facilitates knowledge spillovers among firms, denoted Marshall–Arrow–Romer externalities. Jacobs (1969) instead suggests that diversified urban areas promote the exchange of complementary knowledge among firms, thus generating new ideas and technologies.

In the context of cost structure, localization economies exist when long-run average production costs of firms in a particular industry shift down as the total output of the industry expands. This means economies external to individual firms are transformed into internal scale economies by aggregation to the industry level. Urbanization economies exist when firms' costs shift down due to market and technological spillovers from urban diversity in a densely populated area (Jacobs, 1969).[25] These urbanization economies are again external to individual firms and industries but internal to the urban area as a whole.

Various types of interactions among firms may generate such economies. For example, for urbanization economies, cost advantages from the concentration of firms may simply involve transportation cost savings for interrelated firms that are in close proximity.[26] Goldstein and Gronberg (1984, p. 92) note that smaller firms in large cities may be able to exploit specialized services in large urban areas, so they can carry out their business without having to own all the necessary production tools. Rosenthal and Strange (2004) emphasize the role of urban diversity in terms of labor supply complementarities that can reduce the risks generated by economic fluctuations.

Externalities underlying urbanization and localization economies are often referred to as demand and supply externalities among firms in separate but related industries. That is, 'downstream' firms that purchase intermediate inputs (demanders) would be expected to locate close to 'upstream' firms (suppliers) to save transportation costs and facilitate other interactions. The agglomeration of upstream firms (in the same or similar industries) may increase such externalities for downstream firms. Such interdependencies lead to agglomeration of overall economic activities, such as in Toyota City and the surrounding areas in Aichi Prefecture, Japan, where about 70 percent of intermediate inputs for the automobiles produced were supplied from local automobile-related industries in 2000.

Hirschman (1958) defines supply linkages as forward linkages and demand linkages as backward linkages, and highlights their mutual dependency. Dependencies among firms in the same industry are also sometimes called horizontal linkages while those between industries are called vertical linkages. Myrdal (1957) emphasizes the cumulative causation and synergistic as well as circular nature of such linkages.

The empirical literature on the sources or determinants of agglomeration is as yet quite limited.[27] However, Rosenthal and Strange (2001) provided the basis for a number of subsequent studies. They regress Ellison Glaeser indexes for different zip codes on measures representing Marshall's three sources of localization economies: (1) manufactured and non-manufactured inputs per shipment as indicators of input sharing; (2) innovations per shipment to represent knowledge spillovers (where innovations are defined as the number of new products advertised in trade magazines); and (3) management workers divided by total workers to capture labor pooling. Following their lead, several studies have regressed agglomeration indexes on variables representing agglomeration sources.[28]

For example, Strobl (2004) regresses a location quotient on average plant size (to control for scale economies), labor costs relative to value-added (as a proxy for factor intensity in a Heckscher–Ohlin model), and the value of inputs relative to output (explaining cost and demand linkages), as well as a Ellison and Glaeser co-agglomeration index (to capture spillovers from foreign direct investment – FDI), for Irish manufacturing data. Barrios et al. (2003) regress an Ellison Glaeser index on purchases of goods and services, purchases of energy products, investment in tangible goods, wages and salaries, research and development (R&D) expenditure and average plant size, for Belgium, Ireland and Portugal. Alecke et al. (2006) regress an Ellison Glaeser index on R&D, the share of manufacturing input in total shipments, the shares of workers with specialized occupations and with university degrees, and dummy variables for high-tech and medium-tech industries, for German manufacturing industries. Holmes (1999) tests input sharing between clusters of intermediate input producers when downstream industries are localized by evaluating the relationship between industry localization and vertical disintegration, although he notes that establishing causality between agglomeration and disintegration is ambiguous with cross-sectional data.

Although most studies of agglomeration economies do not address the interaction or endogeneity between the agglomeration of economic activity and knowledge spillovers, Koo (2005, 2007) constructs simultaneous equations that explain both agglomeration and spillovers. His agglomeration variable is the regional relative to national employment density of industry I, and his spillover variable is based on R&D flows between industries.

Effects of agglomeration

As discussed also in Chapter 6, production functions are widely used to measure agglomeration effects, which involve lower costs (economies) or higher productivity for agglomerated firms. In such a framework agglomeration factors are typically assumed to be neutral shift factors (Hicks neutral). For example, the production function of a firm in industry i, region j, may be expressed as:

$$q_{ij} = h(m_{ij}; v_{ij}), \text{ where} \tag{16.40a}$$

$$v_{ij} = g(A_{ij}) f(k_{ij}, l_{ij}), \tag{16.40b}$$

g is a shift factor for the value-added production function h, A_{ij} is a vector of agglomeration variables, q_{ij} is output, v_{ij} is value-added, and k_{ij}, l_{ij} and m_{ij} are capital, labor and intermediate inputs, respectively. For empirical application a Cobb–Douglas (log-linear) form is typically assumed for the production function and the model is often aggregated to the industry level.

A problem with empirical implementation of such a model, however, is measuring the agglomeration factors that appear as components of the A_{ij} vector, such as indicators of localization and urbanization economies. Localization economies are sometimes measured by employment or establishments of the own industry in the region, or a concentration index. Industry output or value-added is sometimes considered a more comprehensive localization measure than employment or establishments, which at the industry level should accommodate increasing returns to scale due to internalization of localization economies.[29] In turn, population level or density is often used to represent urbanization economies, and regional demand size is often represented by transactions estimated from regional input–output tables.

A number of studies have empirically estimated the effects of agglomeration economies in terms of production and productivity.[30] For example, in the recent literature, Henderson (2003) estimates plant-level production functions using plant-level data for machinery and high-tech industries to test localization and urbanization effects.[31] He relies on an urbanization variable similar to the index in equation (16.14) and plant characteristics contained in his panel data to estimate the lagged effects of localized externalities.

Feser (2001, 2002) and Lall et al. (2004) also use plant-level data (for the US and India, respectively), and a translog production function in which agglomeration factors are not Hicks neutral, to estimate agglomeration effects. Feser (2001) focuses on two industries at the three-digit level, where urbanization is measured as total population and localization as total employment in the industry within a 50-mile commuting radius of the plant. Feser (2002) uses the same data to estimate Marshall's three agglomeration sources, where input sharing is measured as potential intermediate supply in the plant's region, labor pooling is measured by the sum of location coefficients, and knowledge spillovers are measured by university research expenditures in the county, all weighted by distance from the plant. Lall et al. (2004) distinguish three sources of agglomeration economies: scale economies from increased market access (measured by transportation network and population weighted access); industry concentration (measured by location quotients); and urban density (measured by available urban utilities and services).

Alternatively, labor productivity is often directly regressed on agglomeration variables. For example, Rigby and Essletzbichler (2002) regress labor productivity on several variables including agglomeration variables to examine productivity differentials across US metropolitan areas. Because they find it difficult to interpret results based on proxies for urban and localization economies such as urban population and industry employment, they construct indexes more directly based on Marshall's externality definitions. In particular, for input sharing they use the Hirschman–Myrdal factor

$$\sqrt{\left(\sum_{i=1}^{I}\omega_{ik}\rho_i\right)\left(\sum_{i=1}^{I}\omega_{ki}\rho_i\right)},$$

where ω_{ik} is the weight of industry i as a supplier of industry k estimated from an input–output table,[32] and for labor pooling they use a sum of squared deviations of the occupational distribution of employment.[33]

Agglomeration effects are sometimes estimated via a labor demand function derived from the first-order condition for profit maximization behavior. For example, Viladecans-Marsal (2004) estimates a labor demand function derived from a CES (constant elasticity of substitution) production function for Spanish cities due to a lack of capital data. Her urbanization economy variables are city population, the squared population as diseconomies caused by excessive city size, industrial employment per capita and a diversification index defined as

$$1 - 0.5\sum_{k\neq i}\left|x_{kj}^{S} - x_{i^*}\right|$$

(similarly to equation 16.14 above). She also uses Moran's I index to evaluate spillover effects of agglomeration economies beyond administrative borders.

Labor demand may also be implied by wages, as in Hanink's (2006) investigation of wage differentials in New England counties for several sectors including manufacturing. He measures urbanization and localization by population and location quotients, and uses Moran's I index to capture the spatial effects of the agglomeration of economic activity. Although his estimation equation is not directly derived from a production function, the log-linear form provides returns-to-scale parameters implying agglomeration effects.

Glaeser et al. (1992) also focus on labor, in terms of employment, due to the lack of output and capital data at the city level in the US. Although there is a potential but sersious problem of omitted-variable bias in such a specification, several papers have applied their framework for different countries using various agglomeration measures. For example, Paci and Usai (2006), using Italian data, regress an average employment growth rate on an employment specialization index, an inverse Herfindahl index for sectoral employment, a Herfindahl index for employee distribution over plants, and average firm size measured by employees per plant (as well as labor supply variables including population and labor force density, and human and social capital). Henderson et al. (1995) apply such a model to eight manufacturing sectors in the US metropolitan areas, including as explanatory variables own-industry employment and concentration (a Hirschman–Herfindahl index of urban diversity). Mano and Otsuka (2000) estimate a regional employment growth equation for Japan using variables similar to Henderson et al. Combes (2000) estimates such a relationship without the wage-change term, based on

a specialization index ($LQ_{ij}^{S(E)}$ in equation 16.18), a diversity index (DIV_{ij}^{C} in equation 16.37), internal economies of scale ($LQ_{ij}^{S(P)}$ in equation 16.17), and a competition indicator (the inverse of a Hirchman–Herfindahl index). Lim (2007) focuses on high-tech industries in US metropolitan areas, and adds to the usual explanatory variables a measure of spatial knowledge diffusion through R&D intensity, proxied by the number of patents per workers discounted by the distance between regions.[34]

Further, Van Oort (2007) extends such specifications by using spatial lag estimates of dependent and independent variables to capture explicitly the spatial structure of proximity. Van Soest et al. (2006) employ a somewhat different index – the lack of industrial diversity defined as the employment share of the five largest industries except industry i – to estimate agglomeration effects on employment growth and new establishment births for a province in South Holland.[35]

Finally, while the theory of spatial agglomeration in a cost function framework was presented by Goldstein and Gronsberg (1984) and extended by Parr (2004), few empirical studies consider agglomeration effects directly in the context of costs except Cohen and Paul (2003, 2005). Cohen and Paul estimate a cost function for food processing manufacture at the US state level, incorporating agricultural product in own and neighboring states as linkage externalities and specifying three agglomeration factors: food production in neighboring states as spatial spillovers; agricultural output in own and neighboring states as supply-side spillovers; and own-state production as demand-side (urbanization) spillovers. Although the cost function approach is useful because it avoids simultaneous problems in estimation, it is difficult to construct a consistent cost data set at the city or county level, where agglomeration economies would be expected to have more significant effects on productivity or cost efficiency.[36]

16.4 Concluding thoughts

From the late 1970s to about 1990, empirical agglomeration studies primarily focused on the estimation of production functions to explain the roles of urbanization and/or localization economies. However, since then many studies have followed the leads of Ellison and Glaeser (1994, 1997), who measure agglomeration through a concentration index of economic activity, and Glaeser et al. (1992), who estimate a reduced form equation of employment or productivity growth. In the former model, the effect of plant size distribution in an industry is incorporated into the agglomeration index, and this index is regressed on various sources of agglomeration suggested by Marshall. In the latter, the main explanatory variables for employment are regional specialization, regional competitiveness and diversity, which are intended to capture urban dynamic externalities.

Although this literature has significantly expanded our understanding of the role of agglomeration economies in economic performance, we believe a number of additional topics deserve further investigation in subsequent literature on agglomeration economies.

One issue that needs more attention is simultaneity. Most empirical studies identify agglomeration economies in terms of higher productivity of firms or industries, which implies that firms or industries obtain performance benefits by locating in areas where related and/or unrelated industries are already located or clustered. Although if agglomeration economies are important to firms they should influence firms' location decisions, firms locating in agglomerated areas also contribute to agglomeration economies. The resulting potential simultaneity problem between sources of agglomeration and firms'

location decisions suggests the need for studies based on behavioral models to estimate these two effects simultaneously,[37] which may involve the complementary use of a dartboard approach and a production function approach.

Urbanization economies also require further consideration, as most empirical studies focus on sources of agglomeration related to localization economies. Population or population density has often been used as a surrogate for urbanization in studies that do address urbanization economies,[38] which may be a useful indicator but is a catch-all. In particular, backward linkage effects (such as home market effects where concentration of employment from density of economic activity attracts more firms) are not necessarily captured or distinguished via such measures. It will be useful in future research to clarify the role of urbanization in terms of both production and consumption and to identify better variables to represent urbanization economies.

Studies of labor market pooling raise unresolved issues in need of attention because there are two perspectives to such pooling. The first involves benefits to the suppliers of labor: pooling means that workers can more easily find another job if they are discharged, although high unemployment rates in metropolitan areas may also attract workers searching for jobs. The second involves benefits to demanders of labor: variety of industrial composition in a region can help firms that seek specific labor skills. Because both supply and demand are therefore at work in regional labor markets, which implies that diversity of both jobs and industries are important, this should be better recognized in future empirical work on labor pooling.[39]

The intangibility of knowledge spillovers also raises issues deserving of attention because it means that such spillovers must be measured indirectly.[40] Although rapid development of the Internet has made it possible to obtain information more easily, face-to-face communication and thus spatial density of activity continue to be important factors in information transmission. More direct measures that reflect such communication – the basis of knowledge spillovers – will be important to develop, potentially through surveys based on questionnaires in agglomerated areas.

Further, regional and sometimes administrative borders have limited the measurement of concentration or specialization. That is, the relationship between spatial interdependence of activities and the range of information spillovers is important to recognize when evaluating agglomeration economies. Recent empirical studies of agglomeration have begun to incorporate spatial lag variables and spatial econometric models,[41] which should be further developed in future studies of agglomeration and spatial correlation to recognize explicitly the spatial behavior of firms and consumers. This will likely require more explicit micro theory models of the cost and input demand behavior of firms, and purchasing choices of consumers.

Finally, although most agglomeration studies have focused on manufacturing activity, inter-industry agglomeration implies linkages also to other industries such as service industries – both locally or internationally. To capture these patterns and their implications for outsourcing more directly it will be desirable to estimate linkage effects between manufacturing and service industries using regional input–output tables.[42]

Notes

1. More specifically, he states that: 'specialization tends to affect location shifts of comparatively narrowly defined economic sectors with homogeneous input structures, while agglomeration involves movements of

more broadly defined sectors comprising goods with very dissimilar input requirements' (Brülhart, 1998, p. 776). Brakman et al. (2001, Chapter 5) geographically illustrate the difference between concentration, specialization and agglomeration.

2. Lafourcade and Mion (2007) and Arbia (2001) distinguish agglomeration from spatial concentration.
3. This index was first proposed by Herfindahl in his PhD dissertation in 1950. Hirschman applied the index to industrial concentration and showed its usefulness.
4. For example, see Traistaru and Iara (2002), p. 7.
5. For a detailed explanation of Lorenz concentration curve, see Brülhart (2001).
6. A similar index is used by Midelfart-Knarvik et al. (2000) to compare the industrial specialization of EU countries. They compare the share of industry *i* in the production of all regions except region *j* instead of s_{i*}.
7. Mulligan and Schmidt (2005) propose an alternative coefficient of specialization for a region.
8. Krugman (1991) conducted a rough comparison of industrial specialization measures between four US regions and four EU regions. Amiti (1999) examines the geographic concentration of industries in the EU countries using location Gini indexes for both location concentration and regional specialization. Midelfart-Knarvik et al. (2002) also investigate the causes of industrial concentration in the EU countries.
9. McCann (2001), Appendix 4.2 explains estimating regional trade using location quotients.
10. Suppose, for example, that agricultural output in a region is 120, and 20 units are exported to foreign countries. The location coefficient of this industry is 1.2 for this region. At the national level, however, agricultural product is imported, so net exports are negative. The national level specialization coefficient (equation 16.23) will be 0.8 and the modified location quotient 0.96. After some calculation one can see that 125 units are supplied by this region so there is a deficit of 5 (= net import) and imports are 25.
11. Lafourcade and Mion (2007) investigate the spatial distribution of manufacturing, depending on the size of plants, using Italian data.
12. See Ellison and Glaeser (1994) for a detailed explanation and proof.
13. See Alecke et al. (2006), p. 21.
14. In the dartboard model γ_i is interpreted as the correlation of each pair of darts.
15. According to Alecke et al. (2006), this index does not show the relationship between the degree of concentration of the industry group and the relative magnitude of the within- and inter-industry component. Maurel and Sédillot (1999) propose a descriptive method to quantify the relative strength of industry- and group-specific agglomeration.
16. Ellison and Glaeser (1997, p. 890) state that the index is designed to facilitate comparison across industries, as well as regions if plant-level data is available.
17. A rigorous explanation of the *K* function is presented in the Appendix of Marcon and Puech (2003).
18. Other researchers have also relied on this index. Barrios et al. (2003) use the Moran index to investigate their choice of spatial units. Cliff and Ord (1981) note that spatial autocorrelation measures were emphasized in the field of quantitative geography in the 1980s, and more recently in the field of 'new economic geography'.
19. The contribution of regional economic diversity to economic performance has typically been investigated in terms of regional economic stability, as elaborated in the survey article by Dissart (2003).
20. This index, also called Shannon's Diversity Index, measures (bio)diversity of species in Ecology.
21. Brülhart and Traeger (2005) apply the entropy index to measure geographic concentration and state several advantages over conventional measures, including statistical meaning. Mori et al. (2005) develop a new measurement of divergence for industrial localization by applying the entropy concept under strong assumptions, and apply it to Japanese regional data.
22. A number of empirical studies have investigated the effects of agglomeration economies; reviews of this literature include Eberts and McMillen (1999) and Rosenthal and Strange (2004). Less empirical literature has been forthcoming on the sources of agglomeration economies, because of measurement and identification issues, as discussed also in Chapter 6.
23. Fujita and Thisse (2002) theoretically explain the determinants of agglomeration.
24. Although it is not easy to identify empirically the productivity contribution of input sharing, Holmes (1999) attempts to evaluate the connection between firm location and input sharing.
25. Glaeser et al. (1992), Henderson et al. (1995) and Rosenthal and Strange (2003) empirically support Jacob's suggestion that these externalities drive urban growth.
26. This is a traditional Weber location decision problem.
27. Dumais et al. (2002) try to decompose the Ellison Glaeser index into the contributions of plant entry and exit on agglomeration. Holmes and Stevens (2002) also decompose the agglomeration index to identify the impact of establishment scale on agglomeration.
28. In the cross-country context, Amiti (1999) estimated the patterns of specialization measured by the Gini coefficient in the EU countries, using intermediate input intensity, relative factor shares of labor, and employment size per firm as explanatory varaibles. Paluzie et al. (2001) applied this specification to data

on 50 Spanish provinces and 30 industrial sectors for 1979, 1986 and 1992. Haaland et al. (1998) estimate the determinants of geographical concentration in 13 EU countries for 35 manufacturing industries, using concentration indexes represented by equation (16.7) and its absolute version, and employment per value-added, per capita wage, labor productivity, and relative expenditure as explanatory variables. Traistaru et al. (2002) evaluate geographic concentration of manufacturing for Bulgaria, Estonia, Hungary, Romania and Slovenia.

29. The parametric treatment of external economies was originally proposed by Chipman (1970). Nakamura (1985) first applied it to estimate urban agglomeration economies of urbanization and localization separately.
30. Because there are extensive reviews of empirical studies on agglomeration economies by Eberts and McMillen (1999) and Rosenthal and Strange (2004), we focus mainly on papers written since 2000.
31. Krugman (1998) also emphasized the importance of empirically investigating linkage externalities, which have not been addressed by many researchers although Midelfart-Knarvik et al. (2002) and Rigby and Essletzbichler (2002) estimated the productivity effects of linkage externalities by constructing linkage indexes using input–output tables for the EU countries and the US.
32. However, it is not evident how they estimated the weights.
33. Maré and Timmins (2006) similarly estimate agglomeration effects on labor productivity for manufacturing firms in New Zealand, based on the concentration index proposed by Maurel and Sedillot (1999) at the local labor market level, industry size at location level, a location quotient (for localization economies) at the industry-location level, and the Hirschman–Herfindahl index at the plant level.
34. Some authors have also estimated such a specification based on total factor rather than labor productivity. For example, Dekle (2002) calculates total factor productivity by one-digit industrial classification level for 47 prefectures in Japan, and regresses them on agglomeration indexes analogous to those used by Glaeser et al. (1992). De Lucio et al. (2002) estimate such a model using data for 50 Spanish provinces and 26 manufacturing industries, with a focus on innovation externalities measured by the provincial value-added growth of specialized industries.
35. Van Stel and Niewenhuijsen (2004) proposed a similar index for diversity before Van Soest et al. (2006).
36. Gao (2004) similarly estimates a simple version of a profit function, which essentially becomes a first-order revenue function where all parameters are linear and input prices are omitted due to the lack of data, for 32 two-digit industries and 29 provinces of China from 1985 to 1993. He includes FDI and exports as well as agglomeration variables such as location quotients as explanatory variables.
37. Koo (2005, 2007) simultaneously estimates the relationship between knowledge spillovers and firms' agglomeration.
38. Based on the model by Ciccone and Hall (1996), Ciccone (2002) estimates the relationship between employment density and labor productivity using 628 NUTS 3 data in main EU countries, the estimation carried out being based on two models of spatial agglomeration. One is based on spatial externalities, and the other on non-tradable inputs produced with increasing returns to scale.
39. Rigby and Essletzbichler (2002) use an industrial labor mix index incorporating occupational composition.
40. Henderson (2007) reviews this issue in the conceptual and empirical literature.
41. For recent applications of spatial autocorrelation, see Fingleton (2001) and Fingleton et al. (2005). Anselin (2003) outlines a taxonomy of spatial econometric model specification that consider spatial externalities.
42. As Henderson (2003) stated, the census of manufacturers does not contain such transaction information. It is thus difficult to estimate scale externalities of producer service industries.

References

Aiginger, K., M. Boeheim, K. Gugler, M. Pfaffermayr and Y. Wolfmayr-Schnitzer (1999), 'Specialisation and (geographic) concentration of European manufacturing', Enterprise DG Working Paper No.1, Brussels: European Commission.

Alecke, B., C. Alsleben, F. Scharr and G. Untiedt (2006), 'Are there really high-tech clusters? The geographic concentration of German manufacturing industries and its determinants', *Annals of Regional Science*, **40**, 19–42.

Amiti, M. (1999), 'Specialization patterns in Europe', *Weltwirtschaftliches Archiv*, **135**, 573–93.

Anselin, L. (2003), 'Spatial externalities, spatial multipliers, and spatial econometrics', *International Regional Science Review*, **26**, 153–66.

Arbia, G. (2001), 'The role of spatial effects in the empirical analysis of regional concentration', *Journal of Geographical Systems*, **3**, 271–81.

Audretsch, D.B. and M.P. Feldman (1996), 'R&D spillovers and the geography of innovation and production', *American Economic Review*, **86** (3), 630–40.

Barrios, S., L. Bertinelli, E. Strobl and A.C. Teixeira (2003), 'Agglomeration economies and the location of industries: a comparison of three small European countries', CORE Discussion Paper, No. 67.

Bertinelli, L. and J. Decrop (2005), 'Geographical agglomeration: Ellison and Glaeser's index applied to the case of Belgian manufacturing industry', *Regional Studies*, **39**, 567–83.
Besag, J.E. (1977), 'Comments on Ripley's paper', *Journal of the Royal Statistical Society B*, **39**, 193–5.
Brakman, S., H. Garretsen and C. van Marrewijk (2001), *An Introduction to Geographical Economics*, Cambridge: Cambridge University Press.
Braunerhjelm, P. and B. Borgman (2004), 'Geographic concentration, entrepreneurship and regional growth: evidence from regional data in Sweden, 1975–99', *Regional Studies*, **38**, 929–47.
Brülhart, M. (1998), 'Economic geography, industry, location, and trade: the evidence', *World Economy*, **21**, 775–801.
Brülhart, M. (2001), 'Evolving geographical concentration of European manufacturing industries', *Weltwirtschaftliches Archiv*, **137**, 215–43.
Brülhart, M. and R. Traeger (2005), 'An account of geographic concentration patterns in Europe', *Regional Science and Urban Economics*, **35**, 597–624.
Chipman, J.S. (1970), 'External economies of scale and competitive equilibrium', *Quarterly Journal of Economics*, **84**, 347–85.
Ciccone, A. (2002), 'Agglomeration effects in Europe', *European Economic Review*, **46**, 213–27.
Ciccone, A. and R.E. Hall (1996), 'Productivity and the density of economic activity', *American Economic Review*, **86**, 54–70.
Cliff, A.D. and J.K. Ord (1981), *Spatial Processes: Models and Applications*, London: Pion.
Cohen, J.P. and C.J. Morrison Paul (2003), 'Spatial and supply/demand agglomeration economies: state- and industry-linkages in the US food system', *Empirical Economics*, **28**, 733–51.
Cohen, J.P. and C.J. Morrison Paul (2005), 'Agglomeration economies and industry location decisions: the impacts of spatial and industrial spillovers', *Regional Science and Urban Economics*, **35**, 215–37.
Combes, P. (2000), 'Economic structure and local growth: France, 1984–1993', *Journal of Urban Economics*, **47**, 329–55.
De Lucio, J.J., J.A.Herce and A. Goicolea (2002), 'The effects of externalities on productivity growth in Spanish industry', *Regional Science and Urban Economics*, **32**, 241–58.
Dekle, R. (2002), 'Industrial concentration and regional growth: evidence from the prefectures', *Review of Economics and Statistics*, **84**, 310–15.
Dissart, J.C. (2003), 'Regional economic diversity and regional economic stability: research results and agenda', *International Regional Science Review*, **26**, 423–46.
Dumais, G., G. Ellison and E. Glaeser (2002), 'Geographic concentration as a dynamic process', *Review of Economics and Statistics*, **84**, 193–204.
Duranton, G. and H.G. Overman (2002), 'Testing for localisation using micro-geographic data', CEPR Discussion Papers, 3379.
Duranton, G. and H.G. Overman (2005), 'Testing for localization using microgeographic data', *Review of Economic Studies*, **72**, 1077–1106.
Duranton, G. and D. Puga (2000), 'Diversity and specialization in cities: why, where and when does it matter?', *Urban Studies*, **37**, 533–55.
Eberts, R.W. and D.P. McMillen (1999), 'Agglomeration economies and urban public infrastructure', in P. Cheshire and E.S. Mills (eds.), *Handbook of Urban and Regional Economics*, vol. 3, New York: North-Holland, pp. 1455–95.
Ellison, G. and E.L. Glaeser (1994), 'Geographic concentration in US manufacturing industries: a dartboard approach', Working Paper No. 3840, NBER.
Ellison, G. and E.L. Glaeser (1997), 'Geographic concentration in US manufacturing industries: a dartboard approach', *Journal of Political Economy*, **105**, 889–927.
Feser, E.J. (2001), 'A flexible test for agglomeration economies in two US manufacturing industries', *Regional Science and Urban Economics*, **31**, 1–19.
Feser, E.J. (2002), 'Tracing the sources of local external economies', *Urban Studies*, **39**, 2485–2506.
Fingleton, B. (2001), 'Theoretical economic geography and spatial econometrics: dynamic perspectives', *Journal of Economic Geography*, **1**, 201–25.
Fingleton, B., D. Igliori and B. Moore (2005), 'Cluster dynamics: new evidence and projections for computing service in Great Britain', *Journal of Regional Science*, **45**, 283–331.
Fujita, M. and J.F. Thisse (2002), *Economics of Agglomeration*, Cambridge: Cambridge University Press.
Gao, T. (2004), 'Regional industrial growth: evidence from Chinese industries', *Regional Science and Urban Economics*, **34**, 101–24.
Gibbs, J. and W. Martin (1962), 'Urbanization, technology, and the division of labour: international patterns', *American Sociological Review*, **27**, 667–77.
Glaeser, E.L, H.D. Kallal, J.A. Scheinkman and A. Shleifer (1992), 'Growth in cities', *Journal of Political Economy*, **100**, 1126–52.
Goldstein, G. and T. Gronberg (1984), 'Economies of scope and economies and agglomeration', *Journal of Urban Economics*, **16**, 91–104.

Haaland, J.I, H.J. Kind, K.H. Midelfart-Knarvik and J. Torstensson (1998), 'What determines the economic geography of Europe?', Discussion Paper No. 19/98, Norwegian School of Economics and Business Administration.
Hanink, D.M. (2006), 'A spatial analysis of sectoral variations in returns to external scale', *Journal of Regional Science*, **46**, 953–68.
Henderson, J.V. (2003), 'Marshall's scale economies', *Journal of Urban Economics*, **53**, 1–28.
Henderson, J.V. (2007), 'Understanding knowledge spillovers', *Regional Science and Urban Economics*, **37**, 497–508.
Henderson, J.V., A. Kuncoro and M. Turner (1995), 'Industrial development in cities', *Journal of Political Economy*, **105**, 1067–90.
Hirschman, A. (1958), *The Strategy of Economic Development*, New Haven, CT: Yale University Press.
Holmes, T.J. (1999), 'Localization of industry and vertical disintegration', *Review of Economics and Statistics*, **81**, 314–25.
Holmes, T.J. and J. Stevens (2002), 'Geographic concentration and establishment scale', *Review of Economics and Statistics*, **84**, 682–90.
Jacobs, J. (1969), *The Economy of Cities*, New York: Vintage.
Koo, J. (2005), 'Agglomeration and spillovers in a simultaneous framework', *Annals of Regional Science*, **39**, 35–47.
Koo, J. (2007), 'Determinants of localized technology spillovers: role of regional and industrial attributes', *Regional Studies*, **41**, 1–17.
Krugman, P. (1991), *Geography and Trade*, Boston, MA: MIT Press.
Krugman, P. (1998), 'What's new about the new economic geography?', *Oxford Review of Economic Policy*, **14**, 7–16.
Lafourcade, M. and G. Mion (2007), 'Concentration, agglomeration and the size of plants', *Regional Science and Urban Economics*, **37**, 46–68.
Lall, S.V., Z. Shalizi and U. Deichmann (2004), 'Agglomeration economies and productivity in Indian industry, *Journal of Development Economics*, **73**, 643–73.
Lim, U. (2007), 'Knowledge externalities, spatial dependence, and metropolitan economic growth in the United States', *Environment and Planning A*, **39**, 771–88.
Malliza, E.E. and S. Ke (1993), 'The influence of economic diversity on unemployment and stability', *Journal of Regional Science*, **33**, 221–34.
Mano, Y. and K. Otsuka (2000), 'Agglomeration economies and geographical concentration on industries: a case study of manufacturing sectors in postwar Japan', *Journal of the Japanese and International Economics*, **14**, 189–203.
Marcon, E. and F. Puech (2003), 'Evaluating the geographic concentration of industries using distance-based methods', *Journal of Economic Geography*, **3**, 409–28.
Maré, D.C. and J. Timmins (2006), 'Geographic concentration and firm productivity', Motu Working Paper 06-08, Motu Economic and Public Policy Research, New Zealand.
Marshall, A. (1920), *Principles of Economics*, London: Macmillan.
Maurel, F and B. Sédillot (1999), 'A measure of the geographic concentration in French manufacturing industries', *Regional Science and Urban Economics*, **29**, 575–604.
McCann, P. (2001), *Urban and Regional Economics*, Oxford: Oxford University Press.
Midelfart-Knarvik, K.H., H.G. Overman, Redding S.J. and A.J. Venables (2000), 'The location of European industry', Economic Papers, No.142, European Commission.
Midelfart-Knarvik, K.H., H.G. Overman, S.J. Redding and A.J. Venables (2002), 'Integration and industrial specialisation in the European Union', *Revue Economique*, **53**(3), 469–81.
Moran, P. (1950), 'A test for serial interdependence of residuals', *Biometrica*, **37**, 178–81.
Mori, T., K. Nishikimi and T.E. Smith (2005), 'A divergent statistic for industrial localization', *Review of Economics and Statistics*, **87**, 635–51.
Mulligan, G.F. and C. Schmidt (2005), 'A note on localization and specialization', *Growth and Change*, **36**, 565–76.
Myrdal, G. (1957), *Economic Theory and Underdeveloped Regions*, London: Duckworth.
Nakamura, R. (1985), 'Agglomeration economies in urban manufacturing industries: a case of Japanese cities', *Journal of Urban Economics*, **17**, 108–24.
Paci, R. and S. Usai (2006), 'Agglomeration economies and growth: the case of Italian local labour systems, 1991–2001', Working Paper 2006/12, Centro Ricerche Economiche Nord Sud, University of Cagliari.
Paluzie, E., J. Pons and D.A. Tirado (2001), 'Regional integration and specialization patterns in Spain', *Regional Studies*, **35**, 285–96.
Parr, J.B. (2004), 'Economies of scope and economies of agglomeration: the Goldstein–Gronberg contribution revisited', *Annals of Regional Science*, **38**, 1–11.
Rigby, D.L. and J. Essletzbichler (2002), 'Agglomeration economies and productivity difference in US cities', *Journal of Economic Geography*, **2**, 407–32.

Ripley, B.D. (1976), 'The second-order analysis of stationary point processes', *Journal of Applied Probability*, **13**, 255–66.
Ripley, B.D. (1977), 'Modelling spatial patterns', *Journal of Royal Statistical Society B*, **39**, 172–212.
Rosenthal, S. and W. Strange (2001), 'The determinants of agglomeration', *Journal of Urban Economics*, **50**, 191–229.
Rosenthal, S.S. and W.C. Strange (2003), 'Geography, industrial organization, and agglomeration', *Review of Economics and Statistics*, **85**, 377–93.
Rosenthal, S.S. and W.C. Strange (2004), 'Evidence on the nature and sources of agglomeration economies', in J.V. Henderson and J.F. Thisse (eds), *Handbook of Urban and Regional Economics*, Vol. 4, New York: North-Holland, pp. 2119–71.
Strobl, E. (2004), 'Trends and determinants of geographic dispersion of Irish manufacturing activity, 1926–1996', *Regional Studies*, **38**, 191–205.
Tokunaga, S. and Y. Akune (2005), 'A measure of the agglomeration in Japanese manufacturing industries using an index by Ellison and Glaeser', (in Japanese), *Studies in Regional Science*, **35**, 155–75.
Traistaru, I. and A. Iara (2002), 'European integration, regional specialization and location of industrial activity in accession countries: data and measurement', Phare ACE Project P98-1117-R, Center for European Integration Studies, University of Bonn, Germany.
Traistaru, I., P. Nijkamp and S. Longhi (2002), 'Regional specialization and concentration of industrial activity in accession countries', Working Paper B16, Center for European Integration Studies, University of Bonn, Germany.
Van Oort, F. (2007), 'Spatial and sectoral composition effects of agglomeration economies in the Netherlands', *Papers in Regional Science*, **86**, 5–30.
Van Soest, D.P., S. Gerking and G.G. Van Oort (2006), 'Spatial impact of agglomeration externalities', *Journal of Regional Science*, **46**, 881–99.
Van Stel, A.J. and H.R. Nieuwenhuijsen (2004), 'Knowledge spillovers and economic growth: an analysis using data of Dutch regions in the period 1987–1995', *Regional Studies*, **38**, 393–407.
Viladecans-Marsal, E. (2004), 'Agglomeration economies and industrial location: city-level evidence', *Journal of Economic Geography*, **4**, 565–82.

17 Measuring the regional divide

Roberto Ezcurra and Andrés Rodríguez-Pose

17.1 Introduction

Measuring the regional divide across countries around the world has become more and more common. The proliferation of subnational data sets with economic information, first in the developed world and later in parts of the developing world, has meant that researchers now have at their disposal a greater array of statistical information to try to understand the dimension and evolution of regional disparities across the world. The increased effort devoted to generating these data sets is also testimony of the growing importance attached by decision-makers to the uneven distribution of wealth within their national boundaries.

Partially as a consequence of this, the number of studies measuring the dimension and evolution of the regional divide throughout the world has increased substantially since the mid-1980s and researchers have progressively improved the techniques aimed at measuring the regional divide. But despite the improvement in information, approaches and methods, it is not entirely clear that our understanding of the dimension of regional disparities and how they evolved is now complete. The numbers of studies that use the same information for similar sets of regions, but that put forward different interpretations of the size, distribution and development of regional disparities, is simply astounding. The case of the European Union (EU) is a clear example. Because of the salience of regional development policies, the EU has become one of the most analysed spaces for regional disparities. As a general rule, these studies rely on the information contained in the REGIO data set by Eurostat, the EU's statistical office, and yet the results differ significantly from one study to another. This is fundamentally because the choice of method of analysis has important implications for our perception of the dimension and evolution of regional disparities in any given space. Different methods analyse different aspects of the variation in a distribution of economic indicators and may thus lead to distinct results and diverse assessments of the dimension and rate of change of regional disparities and to different policy recommendations.

This chapter takes these issues into consideration and aims to provide an overview of the main developments in the measurement of the regional divide, discussing several methodological issues that have arisen since the first attempts to quantify the magnitude of spatial disparities were made. The chapter will highlight the implications of the choice of different methods for our perception of the dimension and evolution of regional disparities and will illustrate these empirically by resorting to the case of the EU-15 during the period 1980–2002.

In order to achieve this aim, the chapter adopts the following structure. First, Section 17.2 presents the traditional approach to the study of regional disparities, based on the analysis of the degree of dispersion in the distribution of regional gross domestic product (GDP) (for example Williamson, 1965; Molle et al., 1980; Terrasi, 1999). We review various measures of dispersion commonly used in the literature to determine the level of

regional inequality and pay particular attention to the criteria taken into consideration when deciding the specific measure used in empirical analyses.

Focusing exclusively on the degree of dispersion could however mask other potentially important features of regional disparities. In particular, as pointed out by Esteban and Ray (1994), the various measures of inequality are inadequate to distinguish whether regions cluster around the average of the distribution or around two or more separate poles. Therefore, from a theoretical point of view, regional inequality may decrease as the degree of polarisation increases (Quah, 1996a). Bearing in mind these considerations, in section 17.3 we examine various measures of polarisation proposed originally in the framework of the literature on personal income distribution. These measures can be readily applied in a spatial context, since the only requisite is to change the unit of analysis.

Section 17.4 reflects the appeal by various authors (for example Quah, 1996a; López-Bazo et al., 1999) to move beyond scalar measures of dispersion and polarisation and to consider the entire distribution under study. It focuses on the non-parametric methodology proposed by Quah (1993, 1996a, 1996b, 1997) in the framework of the economic growth literature to examine the dynamics of the entire cross-sectional distribution.

We include in section 17.5 an illustration of the application of the various approaches considered in our overview to the analysis of regional disparities in the European Union (EU). In this way, we aim to obtain relevant information about a series of methodological issues that arise in the empirical application of the different tools and to assess the potential impact on the results of the method used to measure the regional divide in the European context.

17.2 Regional inequality

The measurement of regional inequality: indices and properties

The vast majority of the studies devoted to the investigation of territorial imbalances have traditionally focused on the analysis of spatial differences. In order to do this, the most straightforward procedure is to calculate some of the numerous measures of dispersion existing in the literature, as a way to summarise in a scalar mode the information on the level of inequality in the distribution. The literature on regional disparities often tends to overlook, however, that the various measures of dispersion are based on different ethical judgements. This issue is especially important, as it implies that each measure of inequality aggregates the information contained in the distribution under consideration in a different way, so that different measures may lead to different results (Sen, 1973). For this reason, it is essential in empirical analyses to check the robustness of the conclusions against various measures of dispersion.

Bearing this in mind, we now offer a brief review of some of the indicators commonly used in the literature when examining the level and evolution of regional inequality. According to this approach, we begin by mentioning two measures that have their origin in descriptive statistics: the coefficient of variation, c, and the standard deviation of the logarithms, v, which can be written as:

$$c = \frac{\sqrt{\sum_{i=1}^{n} p_i (x_i - \mu)^2}}{\mu} \tag{17.1}$$

and

$$v = \sqrt{\sum_{i=1}^{n} p_i (\log x_i - \tilde{\mu})^2} \tag{17.2}$$

where x_i and p_i are respectively the per capita income and the population share of region i in a given year, while $\mu = \sum_{i=1}^{n} p_i x_i$ and $\tilde{\mu} = \sum_{i=1}^{n} p_i \log x_i$.[1] In their unweighted version, these two measures are usually used in the literature to capture the concept of sigma convergence popularised by Barro and Sala-i-Martin (1991, 1992) at the beginning of the 1990s.[2] Recent years have witnessed the publication of numerous papers that apply c and v in this context using regional data (for example Barro and Sala-i-Martin, 1995; Neven and Gouyette, 1995; Rey and Montouri, 1999).

Other authors have instead opted for different indicators of income distribution (for example Molle et al., 1980; Terrasi, 1999; Ezcurra et al., 2005a). Among the measures of inequality employed, the Gini index (G), the generalised entropy class of measures [$GE(\theta)$] and the Atkinson's family of indices [$A(\varepsilon)$] are the most common. These measures can be written respectively as:

$$G = \frac{\sum_{i=1}^{n} \sum_{i=1}^{n} p_i p_j \left| x_i - x_j \right|}{\mu} \tag{17.3}$$

$$GE(\theta) = \begin{cases} \dfrac{1}{\theta(\theta-1)} \sum_{i=1}^{n} p_i \left[\left(\dfrac{x_i}{\mu} \right)^{\theta} - 1 \right] & \theta \neq 0,1 \\ \sum_{i=1}^{n} p_i \log \left(\dfrac{\mu}{x_i} \right) & \theta = 0 \\ \sum_{i=1}^{n} p_i \left(\dfrac{x_i}{\mu} \right) \log \left(\dfrac{x_i}{\mu} \right) & \theta = 1 \end{cases} \tag{17.4}$$

and

$$A(\varepsilon) = \begin{cases} 1 - \left[\sum_{i=1}^{n} p_i \left(\dfrac{x_i}{\mu} \right)^{1-\varepsilon} \right]^{\frac{1}{1-\varepsilon}} & \varepsilon \neq 1 \\ 1 - \exp \left[\sum_{i=1}^{n} p_i \left(\dfrac{x_i}{\mu} \right) \right] = 1 - \prod_{i=1}^{n} \left(\dfrac{x_i}{\mu} \right)^{p_i} & \varepsilon = 1 \end{cases} \tag{17.5}$$

Despite the fact that the detailed analysis of G, $GE(\theta)$ and $A(\varepsilon)$ is beyond the limits of this chapter (see, for example, Sen, 1973, or Cowell, 1995, for further details on this issue), these three indicators present various features that distinguish them from c and v. The Gini index has an appealing geometric interpretation in the context of the Lorenz curve. G is defined as the ratio between the area between the Lorenz curve and the

equidistribution line and the total area under that line. Therefore, the Gini index is equal to the double of the area comprised between the Lorenz curve and the equidistribution line. In turn, the generalised entropy class of measures fulfil various properties that can be applied to carry out different decompositions of overall inequality, which explains the popularity of $GE(\theta)$ in empirical analyses. Finally, the Atkinson's family of indices is derived from a utilitarian social welfare function, and allows to quantify the loss of welfare associated with income dispersion (Atkinson, 1970).

As can be observed, the different inequality measures just defined are all weighted statistics. Accordingly, they take into consideration existing differences in size among the various territorial units employed in the analysis. Although there are some exceptions (see, for example, Gil et al., 2004), this issue has received little attention in the vast majority of the literature on economic convergence, which tends to treat all observations as equal. Nevertheless, as pointed out by Petrakos et al. (2005), the use of a homogeneous weighting scheme may lead to unrealistic results in certain cases, affecting our perception of convergence or divergence trends. The natural alternative is thus to attach a different weight to each observation, reflecting its relative contribution to the sample. Given that our study variable is per capita income, the obvious weights are income or population shares. The use of weighted statistics also has the advantage of reducing the influence of the level of territorial disaggregation considered on the results.

Having reached this point, we now investigate the properties of the measures of dispersion that should be assessed when selecting the series of indicators to be used in empirical analyses. The literature on personal income distribution has highlighted that the main property that should be required in a measure of inequality is the consistency with the Lorenz dominance criterion in the cases when it can be applied. This implies that a distribution x always dominates another distribution x' according to the Lorenz criterion and the measure of dispersion employed should register a greater level of inequality in x' than in x. An inequality measure is consistent with the Lorenz criterion if and only if it fulfils the following properties: the Pigou–Dalton transfer principle, the anonymity principle, independence of scale and independence of population size. The Pigou–Dalton transfer principle implies that any transfer of income from a rich to a poorer region that fails to invert their relative positions must decrease the value of the inequality index. In turn, the anonymity or symmetry principle establishes that the possible permutations of per capita income of the various regions must not modify the level of dispersion, as long as the relative frequencies do not change. The independence of scale implies that the value of the index must remain unaltered regardless of whether the per capita income of all the regions varies in the same proportion or not. Accordingly, this property guarantees that the degree of dispersion registered does not depend on the average per capita income. Finally, the independence of population size establishes that the value of the index must remain unchanged when the number of regions in each level of per capita income is modified in the same proportion.

As can be observed in Table 17.1, the different measures of inequality mentioned above fulfil these basic properties, with the standard deviation of the logarithms being the only exception. Despite its wide use in the literature on economic convergence, v is not consistent with the Lorenz criterion, as it does not satisfy the Pigou–Dalton transfer principle for the whole definition domain of income (Cowell, 1995).

Table 17.1 Basic properties of the measures of inequality

Property	c	v	G	$GE(\theta)$	$A(\varepsilon)$
Pigou–Dalton transfer principle	Yes	No	Yes	Yes	Yes
Anonymity principle	Yes	Yes	Yes	Yes	Yes
Independence of scale	Yes	Yes	Yes	Yes	Yes
Independence of population size	Yes	Yes	Yes	Yes	Yes

Some comments on the decomposition of regional inequality by population subgroups

Although the various measures of dispersion mentioned are useful to describe the level and evolution of regional inequality, they do not provide any information about the origin of regional disparities. In order to study this issue, researchers can resort to different theoretical results that allow them to decompose regional inequality by population subgroups.

This approach requires dividing the set of regions under study into different groups, using some regional characteristic considered relevant for this purpose (for example the country to which a region belongs). This allows the researcher to determine to what extent the observed dispersion can be attributed to the presence of differences between the various groups or within them. This type of decomposition is particularly relevant when it comes to determining the causes of spatial inequality (for example Fan and Casetti, 1994; Terrasi, 1999; Azzoni, 2001).

The fundamentals of this methodology are as follows. Let us suppose that the sample regions have been classified into L exhaustive and mutually exclusive groups. In turn, n_g stands for the number of regions in group g. An inequality measure is said to be additively decomposable if it can be expressed as the sum of a between-group component and a within-group component (Deutsch and Silber, 1999). Specifically, the between-group component is the value of the index of inequality obtained when the per capita income level of all group members coincides with the group average, while the within-group component is equal to the weighted average of internal dispersion in the various groups.

As shown by Shorrocks (1980, 1984) and Foster (1983), the generalised entropy class of inequality measures, $GE(\theta)$, is the only family of indices that are linearly decomposable and that also satisfy the properties commonly required of dispersion measures. In light of this result, $GE(\theta)$ can be written as:

$$GE(\theta) = GE(\theta)^B + \sum_{g=1}^{L} \omega_g GE(\theta)_g \tag{17.6}$$

$GE(\theta)^B$ denotes the generalised entropy class of inequality measures corresponding to $(\mu_1 e_{n_1}, \mu_2 e_{n_2}, \ldots, \mu_L e_{n_L})$, where μ_g is the average per capita income of group g, e_{n_g} is an n_g-dimension vector of ones, and ω_g are the weightings associated with the within-group inequality, which depend exclusively on the population and income shares of each group. In particular, $\omega_g = p_g^{1-\theta} r_g^{\theta}$, where p_g and r_g are respectively the population and income shares of group g. However, $\sum_{g=1}^{L} \omega_g = 1$ only when $\theta = 0,\ 1$. In all other cases the weightings do not add to 1. Therefore, this result makes it advisable to employ $GE(0)$ and $GE(1)$ in empirical analyses. In these two particular cases, expression (17.6) can be written as:

$$GE(0) = \sum_{g=1}^{L} p_g \log\left(\frac{\mu}{\mu_g}\right) + \sum_{g=1}^{L} p_g \left[\sum_{i \epsilon g} p_i \log\left(\frac{\mu_g}{x_i}\right)\right] \tag{17.7}$$

and

$$GE(1) = \sum_{g=1}^{L} p_g \left(\frac{\mu_g}{\mu}\right) \log\left(\frac{\mu_g}{\mu}\right) + \sum_{g=1}^{L} r_g \left[\sum_{i \epsilon g} p_i \left(\frac{x_i}{\mu_g}\right) \log\left(\frac{x_i}{\mu_g}\right)\right] \tag{17.8}$$

17.3 Regional polarisation

The various indices of inequality presented represent a useful way of quantifying the degree of dispersion of the distribution object of analysis. Nevertheless, under certain circumstances, a reduction in the level of regional inequality can be compatible with the presence of greater polarisation, implying the formation of several internally homogeneous regional clusters (Quah, 1996a).[3] This is due to the fact that the conventional measures of inequality usually considered in the literature are incapable of distinguishing whether the observations in the distribution are clustered around the average or around two or more separate poles (Esteban and Ray, 1994; Wolfson, 1994).

Let us consider the following example. Imagine that we have data on the evolution of the regional distribution of per capita income in a given country. We assume that the country in question is undergoing a double process of regional convergence, in which gaps are closing between regions with above-average per capita income at one end of the scale, and between those with below-average per capita income at the other. The process will culminate in a situation in which there will be two homogeneous groups of regions: one of rich and one of poor regions. In this context, however, any measure of inequality that satisfies the Pigou–Dalton transfer principle will register a reduction in overall inequality within the country in question, despite the increasing fracture in the regional distribution of per capita income. This example highlights the need to complement conventional analyses of the degree of dispersion with additional information about the level of polarisation involved.

Since the mid-1990s, several measures of polarisation have been proposed (for example Esteban and Ray, 1994; Wolfson, 1994; Esteban et al., 2007). These measures of polarisation can be readily applied to any spatial context, but the regional analyses that have employed these measures to date are relatively scarce. Some exceptions are Esteban (1994), Le Gallo (2004) and Ezcurra et al. (2006, 2007).

The quantification of regional polarisation

We now present the measures of polarisation proposed by Esteban and Ray (1994) and Esteban et al. (2007). This approach presents various advantages compared with alternative measures of polarisation considered in the literature. Specifically, this approximation is the only one that explicitly incorporates the error generated when partitioning the original distribution into various groups and includes, as a particular case, the index of bipolarisation (polarisation into two groups) proposed by Wolfson (1994).

According to Esteban and Ray (1994), the degree of polarisation of a distribution f into a given number of groups can be measured by means of the following expression:

$$P^{ER}(f,\alpha,\rho) = \sum_{j=1}^{m} \sum_{k=1}^{m} p_j^{1+\alpha} p_k \left|\mu_j - \mu_k\right| \tag{17.9}$$

where μ_j and p_j are, respectively, the average per capita income normalized according to the sample mean and the population share of group j. In turn, $\alpha \in [1, 1.6]$ is a parameter that reflects the degree of sensitivity to polarisation. This measure is very similar to the Gini index. However, the fact that in expression (17.9) p_j is raised to $(1 + \alpha)$ means that this measure of polarisation does not, in theory, follow the same pattern. Before applying this measure, it is necessary to obtain a simplified representation of the original distribution in a series of exhaustive and mutually exclusive groups. This grouping generates some loss of information, depending on the degree of income dispersion in each of the various groups considered. In particular, the generalised measure of polarisation proposed by Esteban et al. (2007) is obtained after correcting the P^{ER} index applied to the simplified representation of the original distribution with a measure of the grouping error.

However, when dealing with personal or spatial income distributions, there are no unanimous criteria for establishing the precise demarcation between different groupings. In order to address this problem, Esteban et al. (2007) apply the algorithm proposed by Davies and Shorrocks (1989) as a way to find the optimal partition of the distribution, ρ^*, that minimises the error term. This means selecting the partition that minimises the Gini index value of within-group inequality, $G(f) - G(\rho^*)$.[4] The measure of generalised polarisation proposed by Esteban et al. (2007) thus becomes:

$$P^{EGR}(f,\alpha,\rho^*,\beta) = P^{ER}(f,\alpha,\rho^*) - \beta[G(f) - G(\rho^*)] \tag{17.10}$$

where $\beta \geq 0$ is the weighting parameter for the error term in expression (17.10).

P^{EGR} includes the index of bipolarisation derived by Wolfson (1994) as a particular case. Wolfson (1994) measures the level of polarisation associated with a partition of the original distribution into two groups divided by the median using the following expression:

$$P^{W}(f) = 2\frac{\mu}{m}[1 - 2L(0.5) - G(f)] \tag{17.11}$$

where m and $L(0.5)$ are, respectively, the median and the value of the Lorenz curve at the 50th percentile of the f distribution. The Wolfson's (1994) polarisation measure does, however, not minimise the error arising from the partition of the distribution into two groups. In any event, Esteban et al. (2007) show that both indices satisfy the following relationship:

$$P^{EGR}(f,\alpha = 1,\rho',\beta = 1) = \frac{m}{2}P^{W}(f) \tag{17.12}$$

where ρ' denotes the partition of the original distribution into two equally sized groups.

Exploring the origin of regional polarisation

Other regional characteristics besides per capita income may also be relevant for polarisation. Unfortunately, the P^{EGR} measure is not useful in principle to examine the degree of regional polarisation between some exogenously given clusters defined independently of regional per capita income.

As pointed out by Gradín (2000) in the context of the literature on personal income distribution, a first possibility to study this issue is to adapt the generalised measure of

polarisation introduced by Esteban et al. (2007) to this new framework. Gradín (2000) derives different measures of polarisation based on P^{EGR}, incorporating different characteristics of the households, such as education level or the position occupied in the labour market, into the analysis. This method has been applied by Ezcurra et al. (2005b) when examining the impact of the national component, geographical location, and sectoral composition of economic activity on polarisation across EU regions.

Zhang and Kanbur (2001) have, in turn, developed an alternative approach to analyse this issue based on the decomposition of aggregate inequality by population subgroups. These authors propose the use of a measure of polarisation defined as the ratio of between-group inequality to within-group inequality. Their index of polarisation takes the following form:

$$P^{ZK}(f) = \frac{GE(\theta)^B}{\sum_{g=1}^{L} \omega_g GE(\theta)_g} \tag{17.13}$$

In expression (17.13) the between-group component, $GE(\theta)^B$, provides information on the level of dispersion of the average per capita incomes of the various groups considered in the analysis, while the within-group component, $\sum_{g=1}^{L} \omega_g GE(\theta)_g$, indicates the internal cohesion within them. As can be observed, $P^{ZK}$is a measure of polarisation that captures the average distance between the groups in relation to the per capita income differences observed within them.

17.4 The distributional approach to measuring the regional divide

The analysis of regional disparities goes beyond, however, the calculation of a set of indices of inequality and polarisation. In fact, inequality and polarisation statistics do not provide an accurate description of the entire distribution. In order to address this issue, this section presents the non-parametric approach proposed by Quah (1993, 1996a, 1996b, 1997) as a means to examine the evolution over time of the entire cross-sectional distribution, putting particular emphasis on both the change in its external shape and on the intra-distribution dynamics. This approach arose as a result of attempts to overcome some of the limitations of the conventional convergence analyses based on the estimation of cross-sectional growth regressions (see Magrini, 2004, for further details). Nevertheless, this method can also be applied to the study of regional disparities (for example Quah, 1996c; López-Bazo et al., 1999; Magrini, 1999).

The external shape of the distribution

Economic data are often asymmetric or present multiple modes. These potentially interesting data features cannot, however, be captured by the various indices introduced in the preceding sections. For this reason, the external shape of the distribution being analysed deserves greater attention. The literature tends to examine this issue by means of non-parametric estimation techniques, which avoid the need to specify a particular functional form beforehand. There are several major advantages to using this type of approximation, given the lack of generality and flexibility associated with parametric approaches.

The use of a non-parametric methodology implies the need to select a method to smooth the data. A first option is to employ histograms, the oldest and best-known

non-parametric density function estimator. Histograms are useful to describe certain data features, but present several drawbacks that make them unsuitable for the estimation of the density function of the distribution under study, $f(x)$.[5] Other possible approximations include, for example, the naive estimator, the nearest-neighbour method, the orthogonal series estimator, or the variable kernel method. However, the most popular alternative is to use kernel smoothing. According to this approach, the density function of the distribution considered can be estimated from:

$$f(x) = \frac{1}{h}\sum_{i=1}^{n} p_i K\left(\frac{x - x_i}{h}\right) \tag{17.14}$$

where K is a kernel function that integrates to 1 and h is the smoothing parameter. Following the strategy adopted in the preceding pages, we have introduced the population share of each region into the analysis, thus taking into consideration the different size of the various territorial units.[6]

The use of expression (17.14) requires the choice beforehand of a kernel function and a smoothing parameter. The statistical properties of the proposed estimator depend on this double choice. When it comes to assess the relevance of the kernel function in this context, Marron and Nolan (1988) have shown that it is possible to rescale K, so that there is practically no difference between the estimates obtained from different kernel functions. The choice of the smoothing parameter is however much more important, as it determines the amplitude of the bumps. Choosing an excessively large h generates a reduced number of bumps (oversmoothing), which can contribute to conceal several features that might be present in the data, such as multimodal structures. On the contrary, the selection of a smaller h gives rise to an excessive number of bumps (undersmoothing), which could make the observation of the true structure of the data difficult.

Intra-distribution mobility

The methodology described helps to overcome some of the limitations of the measures of inequality and polarisation introduced in the preceding sections by considering the entire distribution. Nevertheless, the estimation of different density functions is based exclusively on the information provided by a series of cross-sectional observations of the distribution being analysed. Consequently, as in the case of measures of inequality and polarisation, this approximation ignores the fact that various economies may modify their relative positions over time. This issue is particularly relevant when assessing from a normative point of view the evolution of regional disparities. In order to overcome this drawback, we now proceed to complete our previous results by addressing the issue of intra-distribution mobility.

Here we follow Quah's methodology (1993, 1996a, 1996b, 1997). According to Quah's approach, the distribution of regional per capita income in period t has an associated probability measure, ϕ_t. Our aim is to describe the law of motion of the stochastic process $\{\phi_t, t \geq 0\}$. The simplest way of modelling the distribution dynamics is using a first-order dependence specification:

$$\phi_t = T^*(\phi_{t-1}, u_t) = T^*_{u_t}(\phi_{t-1}) \tag{17.15}$$

where u_t is a sequence of disturbances, while T^* stands for an operator that maps probability measures in $t-1$ and disturbances in t to probability measures in t. For simplicity, we assume that the disturbances are included in the definition of the operator, $T^*_{u_t}$.

A first way to use equation (17.15) for the study of the distribution dynamics is to make the space of economies discrete. This makes operator $T^*_{u_t}$ a transition probability matrix, M_t. Furthermore, by assuming that the underlying transition mechanism is time-invariant, the model is a time-homogeneous finite Markov chain:

$$\phi_{t+1} = M'\phi_t \tag{17.16}$$

Accordingly, for all $s \geq 1$we have that:

$$\phi_{t+s} = (M^s)'\phi_t \tag{17.17}$$

This approach allows for the estimation of the ergodic distribution that informs on the long-run limit of the distribution analysed. The ergodic distribution corresponds to the limit of equation (17.17) as $s \to \infty$:

$$\phi_\infty = M'\phi_\infty \tag{17.18}$$

There are a series of works that apply this methodology to the study of regional disparities, using the information provided by different transition probability matrices (for example López-Bazo et al., 1999; Le Gallo, 2004). This discrete approximation allows the researcher to calculate some of the measures proposed in the literature on personal income distribution to quantify the degree of mobility of the different transition probability matrices (Pekkala, 2000). However, the findings reached through this approach may be sensitive to the criterion used to define the transition probability matrix in each case. As there is no procedure for determining the optimum number of states and the boundaries between them, an arbitrary decision is taken. Hence, as pointed out by Bulli (2001), different methods may provide different results and even affect the Markov properties of the dynamic process. These problems can be overcome by considering that operator $T^*_{u_t}$ in equation (17.15) is a stochastic kernel. The stochastic kernel corresponds to a continuous version of a transition probability matrix and is obtained by estimating the density function of the distribution in a given period $t+s$, conditioned on the values of a previous period t. More specifically, the joint density function at moments t and $t+s$ is estimated by the kernel method and then divided by the implicit marginal distribution at t in order to obtain the corresponding conditional probabilities (see Durlauf and Quah, 1999 for further technical details). Due to the methodological problems raised by the employment of transition probability matrices in this context, the use of stochastic kernels in regional studies has increased considerably during recent years (for example Johnson, 2000; Overman and Puga, 2002; Ezcurra et al., 2005a).

17.5 Empirical application: regional disparities in the EU

This section assesses empirically the usefulness of the various methodological approaches described in the previous pages when analysing the regional divide existing in any given geographical area. We use EU regions as our example, with the aim of testing a series of

methodological issues that arise in the empirical application of the different methods examined.

The last 15 years have witnessed the publication of a large amount of research on spatial disparities in the EU (see Eckey and Türk, 2006, for a review of this literature). There are various reasons that explain the rising interest in this issue. High among them come the major advances made since the mid-1980s in economic growth theory (Barro and Sala-i-Martin, 1995), and the development of the 'new economic geography' models (Ottaviano and Puga, 1998). But beyond the academic debate, the growing relevance of this topic is largely due to the strong focus placed on achieving economic and social cohesion since the mid-1980s by the European Union (European Commission, 2004).

The perception of the regional divide in the EU is not only affected by which methodology is used in order to assess regional disparities. The choice of spatial units for the analysis of territorial imbalances within the EU also influences results. Most researchers working on European regional inequality have resorted to the so-called NUTS as the unit of analysis. NUTS is the French acronym for 'Nomenclature of Territorial Units for Statistics', a hierarchical classification of subnational spatial units established by Eurostat according to administrative criteria. In this classification, NUTS-0 corresponds to the country level, while increasing numbers indicate increasing levels of disaggregation. Our analysis focuses on NUTS-2 regions. This is due to various reasons. First, NUTS-2 is the territorial unit most commonly employed for the examination of regional disparities in Europe. Second, NUTS-2 are particularly relevant in terms of EU regional policy, as this is the spatial level of eligibility for Objective 1 since the 1989 reform of the European Structural Funds. The choice of NUTS-2 is, however, not devoid of problems. NUTS-2 regions differ considerably in size, population and level of autonomy (Rodríguez-Pose, 1998, 1999). For this reason, various authors have proposed alternative classifications based on Functional Urban Regions (Cheshire and Carbonaro, 1996; Magrini, 1999), defined on the basis of core cities and taking into account existing commuters flows (see Cheshire and Hay, 1989, for further details). The level of spatial aggregation used in the analysis affects the results as a consequence of the modifiable areal unit problem (MAUP) or with some form of the ecological fallacy problem (Arbia, 1989). Accordingly, particular attention should be paid to this issue by regional researchers, especially when comparing the level of regional disparities in different countries or geographical areas.

The empirical application carried out in this chapter is based on data drawn from the Cambridge Econometrics regional database. The per capita income of 196 NUTS-2 regions of the EU-15 has been calculated for the period 1980–2002 from data on GDP and population. The data provided by Cambridge Econometrics are based mainly on information supplied by REGIO, the Eurostat regional database. Cambridge Econometrics has filled the gaps in the original database, especially with regards to data relating to the late 1970s and early 1980s, using national statistics and interpolation methods.[7]

Inequality and regional polarisation

We begin our analysis by examining the evolution of the level of dispersion of the regional distribution of per capita GDP in the EU over the period 1980–2002. As mentioned in section 17.2, the literature devoted to the study of personal income distribution has highlighted that the results may differ, at times substantially, according to which measure of inequality is used in the analysis. Given the obvious difficulty that arises from the fact that

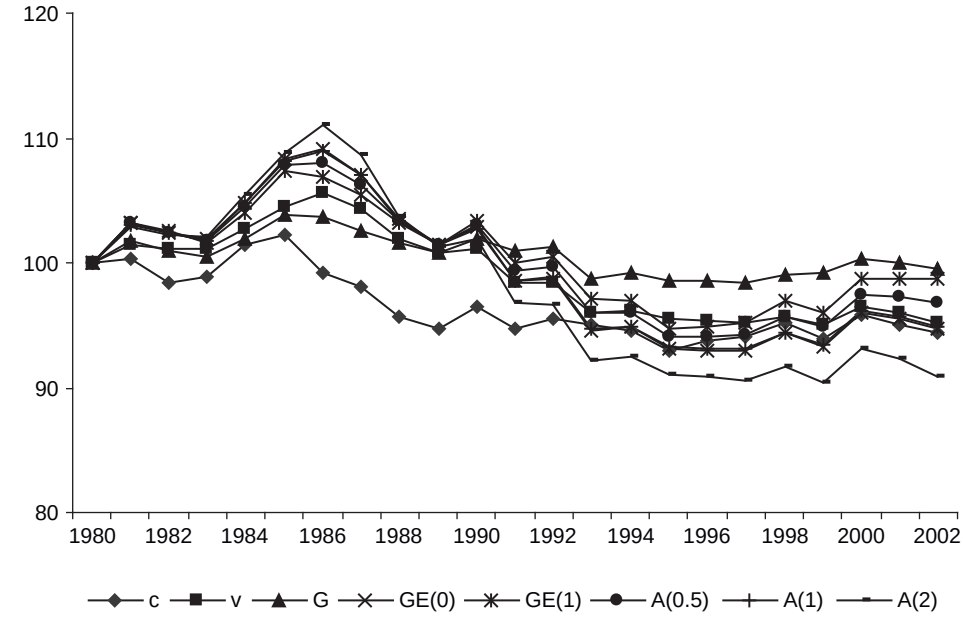


Figure 17.1 Regional inequality in per capita GDP (1980 = 100)

different indicators may give rise to different orders of the distributions compared, the robustness of our results needs to be checked against different measures of dispersion. The degree of regional inequality in the EU-15 is thus examined by means of the information provided by the coefficient of variation, c; the standard deviation of the logarithms, v; the Gini index, G; the two indices popularised by Theil (1967), which correspond to the generalised entropy class of inequality measures when $\theta = 0, 1$, $GE(0)$ and $GE(1)$; and the Atkinson's family of indices with different values of the inequality aversion parameter ε, $A(0.5)$, $A(1)$ and $A(2)$.

Given the large differences in population among European regions,[8] all measures have been calculated including the population shares, which is consistent with the approach commonly adopted in the literature on personal income inequality. This allows for partially reducing the potential problems arising from the degree of territorial disaggregation considered in our study. The use of population-weighted measures is, however, not usual in the studies on regional convergence, which tend in general to assign the same weight to all regions included in the sample.

Figure 17.1 reports the evolution of all the indices of regional inequality in per capita GDP. These indices show that the degree of dispersion of regional per capita GDP decreased between 1980 and 2002. In particular, the values of the various indices fell between 1 and 9 per cent during the study period. Hence, the magnitude of the reduction in the level of dispersion depends on the specific measure considered. Furthermore, our results indicate that the rate of decline in regional inequality was not uniform over time. Three different phases can be distinguished in this context. First, regional inequality rose during the early 1980s, reaching its maximum level around 1985. This upward trend was to change in the next ten years when the main reduction in the level of dispersion took

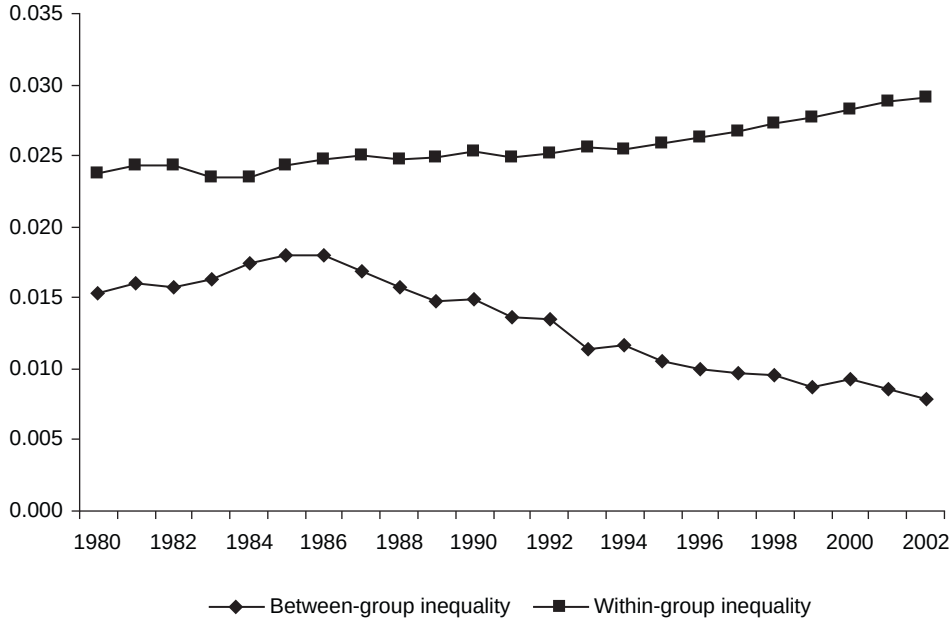


Figure 17.2 Decomposition of regional inequality in per capita GDP

place. Finally, from the mid-1990s onwards, regional inequality did not register significant variations, remaining practically constant.

The use of measures adopted from the literature on personal income inequality allows us to identify some interesting features concerning the evolution over time of the regional distribution of per capita GDP in the EU. For example, the reduction experienced by the Atkinson indices calculated grows as the value of the inequality aversion parameter ε increases. This is particularly relevant, as it suggests that the observed decrease in the degree of dispersion over the period 1980–2002 coincided with an improvement in the relative situation of those regions situated at the lower end of the distribution.

As every region included in the sample belongs to a specific country, it is therefore reasonable to ask whether, for a given level of regional inequality, the existing territorial imbalances are greater between countries or within countries. We therefore apply the theoretical results on additive decomposition of inequality measures that were introduced in section 17.2.

Figure 17.2 presents the decomposition of $GE(0)$ according to expression (17.7).[9] The results reveal the simultaneous presence of between-country convergence and regional divergence in our sample, which is in line with previous findings (Esteban, 1994; Puga, 2002). The continuous decline in the differences between countries in average per capita GDP from the mid-1980s onwards has coincided with a steady rise in the value of the within-group component. It is worth noting that, while at the beginning of the study period, the between-group component of $GE(0)$ explained 40 per cent of the global inequality, by 2002 this percentage had dropped to 21 per cent. Therefore, if within-group disparities were eliminated at the end of the sample period, while keeping between-group dispersion constant, the aggregated inequality would have been reduced by 79 per cent.

This raises a series of potentially major implications for the future design of the EU regional policy (European Commission, 2004). The analysis carried out shows that in 2002 inequality due to the within-group component of $GE(0)$ played a relevant role, making policies aimed at correcting imbalances between countries less likely to have an impact. Under these circumstances a reinforcement of the role of specific redistribution policies within countries would be advisable.

In order to complete these results, Figure 17.3 shows the evolution of regional inequality within the various EU member states considered.[10] Internal regional inequality varies considerably across the sample countries. The highest $GE(0)_g$ values are found in Belgium, Italy, the United Kingdom and Portugal, making them the countries with the highest levels of dispersion in the regional distribution of per capita GDP. At the opposite end of the scale we find Finland, Ireland, Sweden and Germany, which are characterised by the lowest levels of regional inequality during the study period. Caution should be, however, exercised when assessing the implications of Figure 17.3, as the results obtained may be affected by the different number of NUTS-2 regions in each country. In order to explore this issue, we proceeded by calculating the correlation coefficient between the average values of $GE(0)_g$ over the years examined and the number of NUTS-2 regions in each country. The coefficient estimated is 0.289 and lacks statistical significance (p-value = 0.338).

The information provided by Figure 17.3 reveals the absence of a clear evolutionary pattern across EU countries. Between 1980 and 2002 the degree of dispersion of regional per capita GDP rose in Germany, Spain, France, Ireland, Finland, Sweden and the United Kingdom, countries that in 1980 represented 67 per cent of total population in our sample. This explains the evolution experienced by the within-group component of $GE(0)$ discussed above (Figure 17.2). In contrast, the value of $GE(0)_g$ fell in Belgium, Greece, Italy, the Netherlands, Austria and Portugal.[11]

The various dispersion measures considered so far do not allow us to assess whether the sample regions are clustered around the average of the distribution or around two or more separate poles. This implies that the observed reduction in the degree of dispersion in the regional distribution of per capita GDP may be compatible with the presence of polarisation patterns within the EU. In order to investigate this issue in greater depth, we apply the approach proposed by Esteban and Ray (1994) and Esteban et al. (2007). This methodological approximation was described in detail in section 17.3 of this chapter. In order to check the robustness of the results, we have considered different degrees of sensitivity to polarisation in our analysis. Specifically, according to the notation employed previously, $\alpha = 1, 1.3, 1.6$. Following Esteban et al. (2007), $\beta = 1$ in all cases.[12]

Figure 17.4 shows the time path for the generalised measure of polarisation derived by Esteban et al. (2007), when the algorithm proposed by Davies and Shorrocks (1989) is used to define a simplified two-group representation of the original distribution of per capita GDP across the European regions. In the two-group case, the optimal partition of the distribution is characterised by the fact that the per capita GDP level that separates the two groups is equal to the sample average. When our 196 regions are divided into two groups following this criterion, it is possible on average to account for 69 per cent of total dispersion as measured by the Gini index. Hence, the amount of internal inequality left unexplained by the grouping is 31 per cent. As observed in Figure 17.4, Europe has wit-

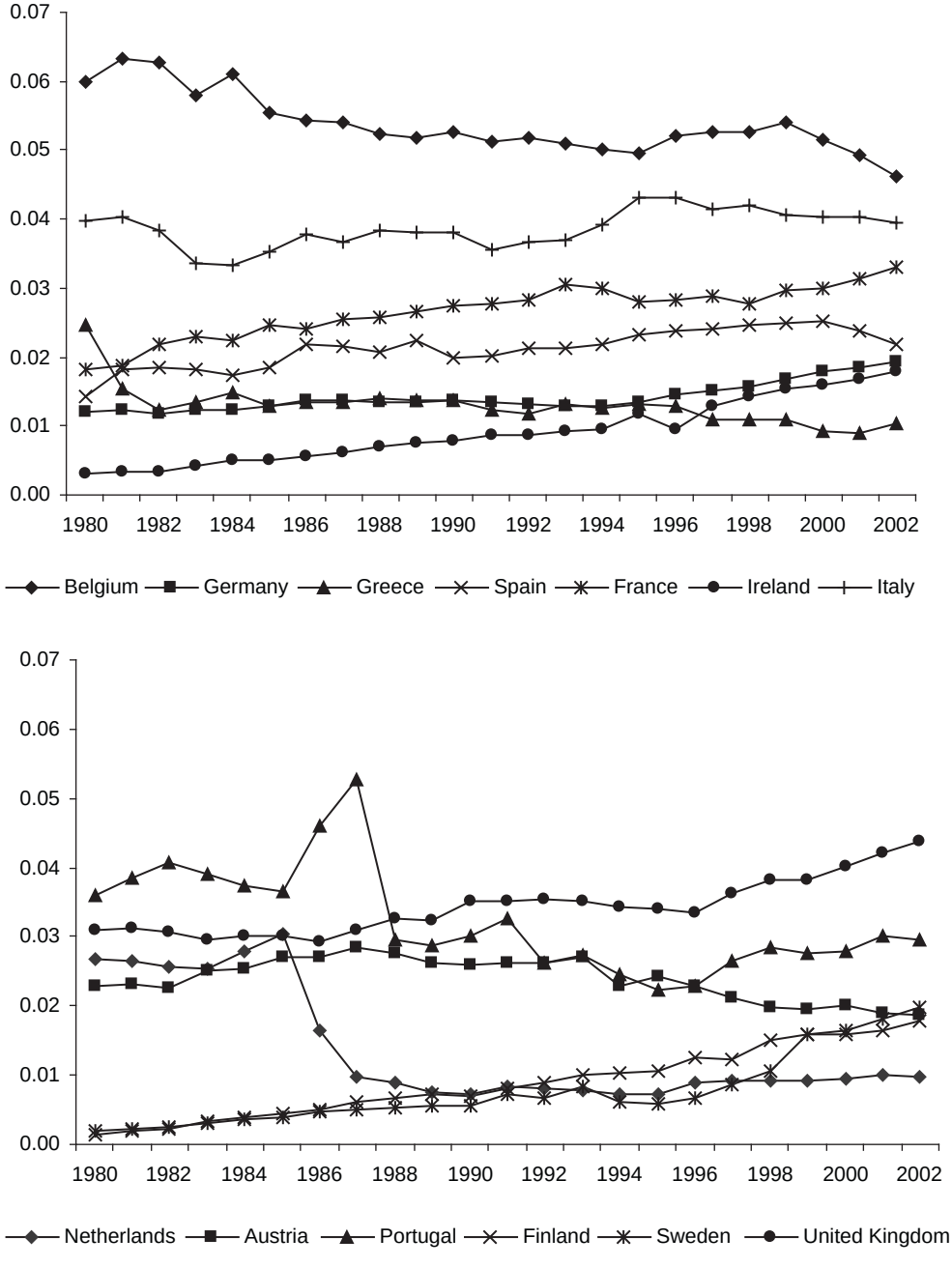


Figure 17.3 Regional inequality in per capita GDP within countries

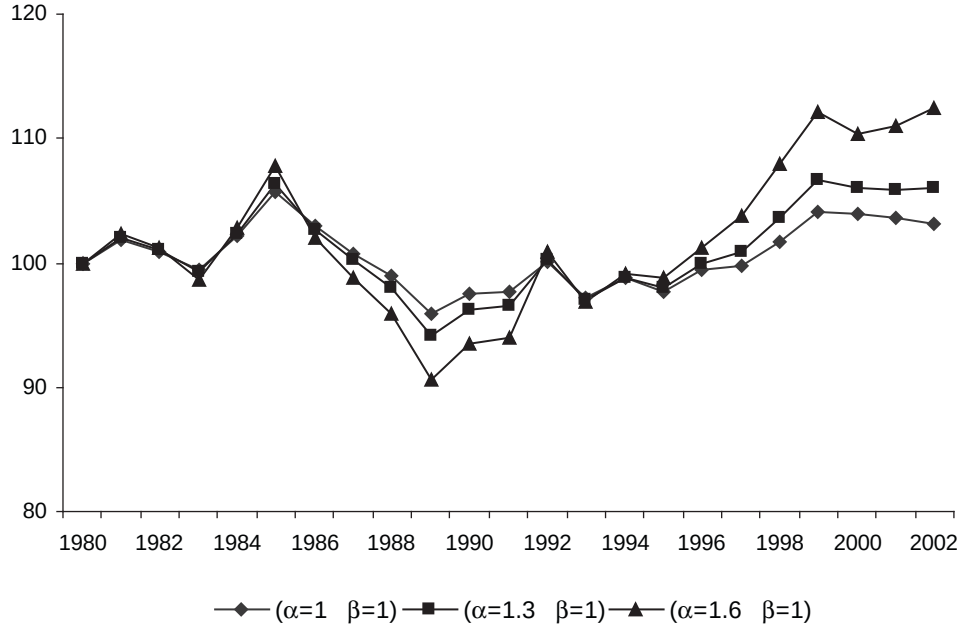


Figure 17.4 Regional polarisation: two groups (1980 = 100)

nessed an increase in the degree of bipolarisation during the period of analysis. The values of P^{EGR} rose by between 3 and 12 per cent between 1980 and 2002, depending on the weight assigned to the sensitivity polarisation parameter. This evolution reveals that the observed decrease in regional inequality coincided with an increase in the level of bipolarisation of the regional distribution of per capita GDP. The evolution of P^{EGR} went through different phases: regional bipolarisation rose during the early 1980s, then decreased towards the end of the decade, reaching its minimum level in 1989. The downward trend was reversed in the 1990s. The value of P^{EGR} increased considerably throughout that decade, coinciding with the advances made in the economic integration process and the increase of the funds devoted to promote economic and social cohesion within the EU. Finally, in what might be interpreted as the beginning of a new phase, from 1999 to 2002, regional bipolarisation remained practically constant.

Using a two-group representation of the original distribution, we can analyse the evolution of regional polarisation without any disproportionate loss of information. We run the risk, however, of interpreting a decrease in polarization, when what is actually taking place is a division of the distribution into three poles (Esteban and Ray, 1994). Likewise, the dualised view of the European economy underlying the above analysis may be an oversimplification. As a first step towards addressing these problems, we now consider an alternative classification of European regions into three groups: one made up of regions with per capita GDP around the EU average, and two extreme groups.

Figure 17.5 reports the evolution of regional polarisation when the original distribution is split into three groups, by applying the algorithm proposed by Davies and Shorrocks (1989). The three-group representation explains on average 85 per cent of global inequality, versus the 69 per cent corresponding to the previous partition. Figure

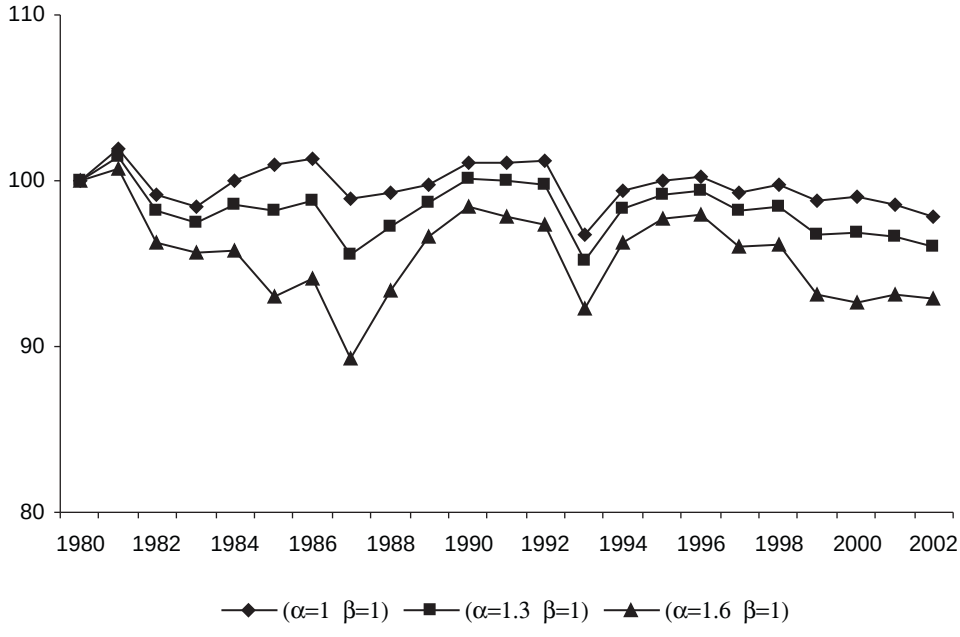


Figure 17.5 Regional polarisation: three groups (1980 = 100)

17.5 shows an overall decline in polarisation during the 1980–2002 period, which contrasts with the evolution of bipolarisaton just commented upon. P^{EGR} values, in particular, fall by between 2 and 7 per cent over the 23 years considered, depending on the value adopted by the α parameter. This was mainly due to the evolution experienced by the generalised measure of polarisation during the 1980s, since P^{EGR} remained practically constant throughout the 1990s.

In short, the information provided by Figures 17.4 and 17.5 indicates that polarisation did not evolve in parallel with regional inequality. Therefore, the empirical evidence presented above suggests that, beyond the conceptual differences existing between the two notions, there are evident discrepancies between polarisation and regional inequality in the EU. This conclusion highlights the need for carrying out separate analyses of these two phenomena in empirical applications, something that has been rare in the literature devoted to the study of regional disparities in different geographical settings.

In order to check whether the changes experienced over time by the various measures of inequality and polarisation calculated above are statistically significant, we consider two alternative approaches proposed for the examination of inferential issues within the framework of the literature on personal income distribution. The first is based on the existing theoretical results for asymptotic distributions (Maasoumi, 1997). The use of this approximation in the regional context is however questionable, as regional samples are usually much smaller than those employed in the literature on personal income distribution. The second alternative is based on bootstrap methods, which can be employed in principle to generate the corresponding confidence intervals (Mills and Zandvakili, 1997; Biewen, 2001). According to this approach, the sampling distribution of the various measures of inequality and polarisation may be estimated by multiple random resam-

pling. This assumption is again difficult to justify in regional studies, as there is a large amount of empirical evidence for different geographical settings that suggests that measurement errors are probably spatially autocorrelated (for example Quah, 1996c; Lopéz-Bazo et al., 1999; Rey and Montouri, 1999). Accordingly, standard bootstrap methods cannot be applied in this context.

The various arguments put forward highlight that inferential methods to test formal hypotheses regarding measures of inequality and polarisation have yet to be developed for the purposes of regional analysis. A step in this direction has been made by Rey (2001), who investigates the possibility of using a new approximation based on random spatial permutations of the actual values of the variable being considered for a given map pattern, in order to test hypotheses about the decomposition of the generalised entropy class of inequality measures into its between-group and within-group components.

The distribution dynamics

The series of measures calculated in this section provide relevant information on the evolution of regional disparities in the EU. However, when interpreting the results obtained so far, the fact that those statistics aggregate all the information contained in the regional distribution of per capita GDP should not be overlooked. The picture of the distribution is thus partial and somewhat inaccurate.

This implies a need to examine also the external shape of the EU regional distribution of per capita GDP over the period 1980–2002. In order to do this we estimate non-parametrically the population-weighted density functions of the distribution according to the procedure described in section 17.4.[13] The results are shown in Figure 17.6. Following common practice in the literature, each region's per capita GDP has been normalised with respect to the EU average (100). All the density functions estimated show that the majority of the regions considered enjoyed a level of development around the European average between 1980 and 2002. However, this does not imply that the initial situation remained stable during these 23 years: the density shifted gradually towards the left throughout the 1990s, indicating that the population share that lived in regions with a per capita GDP below the average increased. Our estimates also highlight that the least favoured areas of the Union succeeded in partially narrowing their per capita GDP gaps in relation to the European average, which is consistent with the results obtained previously by calculating the Atkinson's inequality index (Figure 17.1). Likewise, the density located at the upper end of the distribution did not register outstanding changes over the sample period.

The information provided by Figure 17.6 is based exclusively on a series of cross-sectional observations of the distribution under study and does not, therefore, take into consideration that the different regions may modify their relative positions over time. In order to address this shortcoming and to complete our previous results, we now examine the intra-distribution mobility.

Figure 17.7 depicts the population-weighted stochastic kernel estimated for one-year transitions. The three dimensional graph shows how the cross-sectional distribution at t evolves into that observed at $t+1$.[14] The stochastic kernel gives the probability distribution of per capita GDP at $t+1$ for regions with a given value of the study variable at t. If the probability mass is concentrated around the main diagonal, the intra-distribution dynamics are characterised by a high level of persistence over time and, therefore, by low

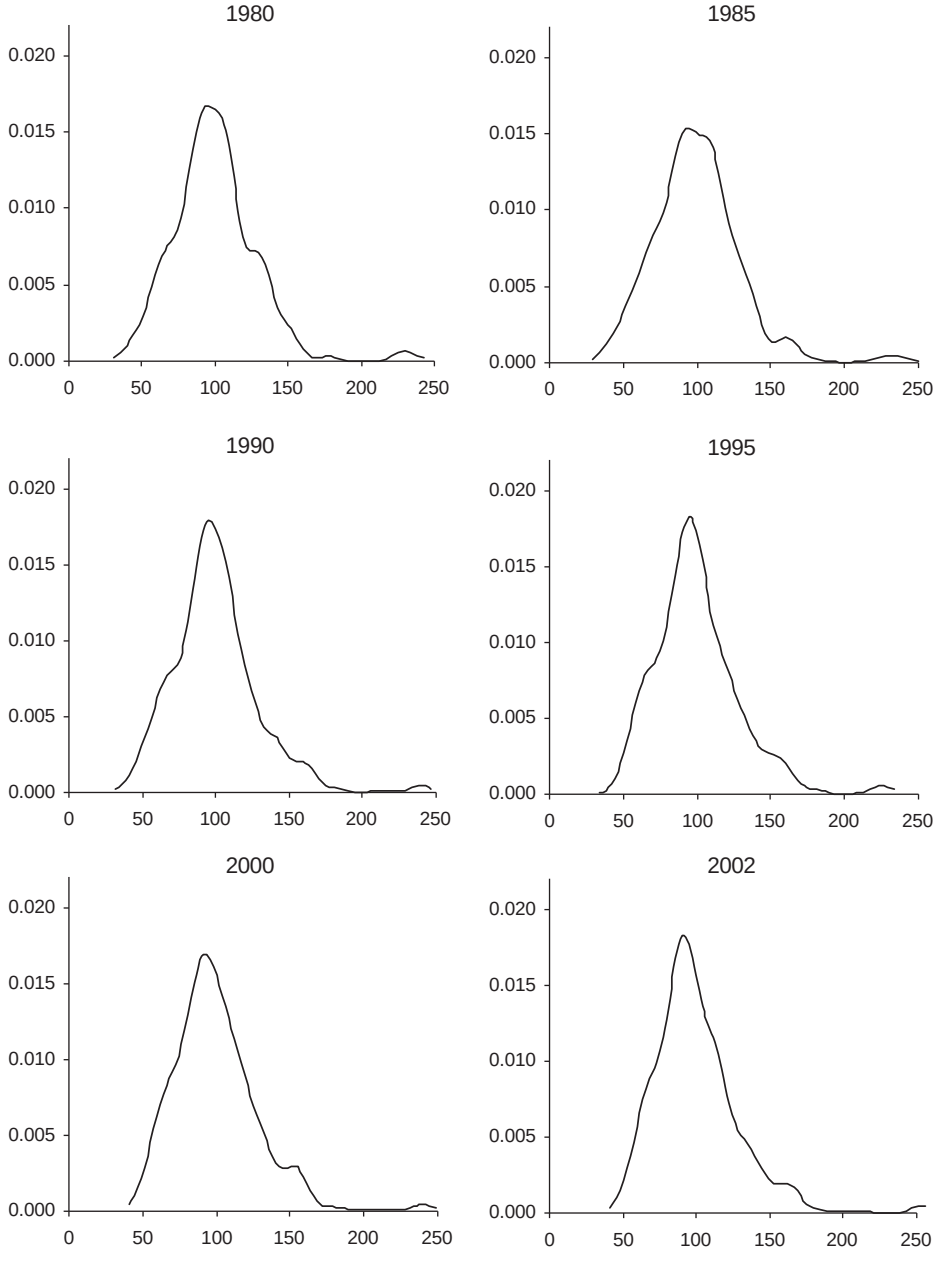


Figure 17.6 Density functions of EU-relative regional per capita GDP distribution

regional mobility. If, on the other hand, the probability mass is located on the opposite diagonal, this indicates that regions situated at both ends of the distribution exchanged their relative positions during the study period. Finally, the probability mass could theoretically accumulate parallel to the *t* axis. This would reflect the presence of a process

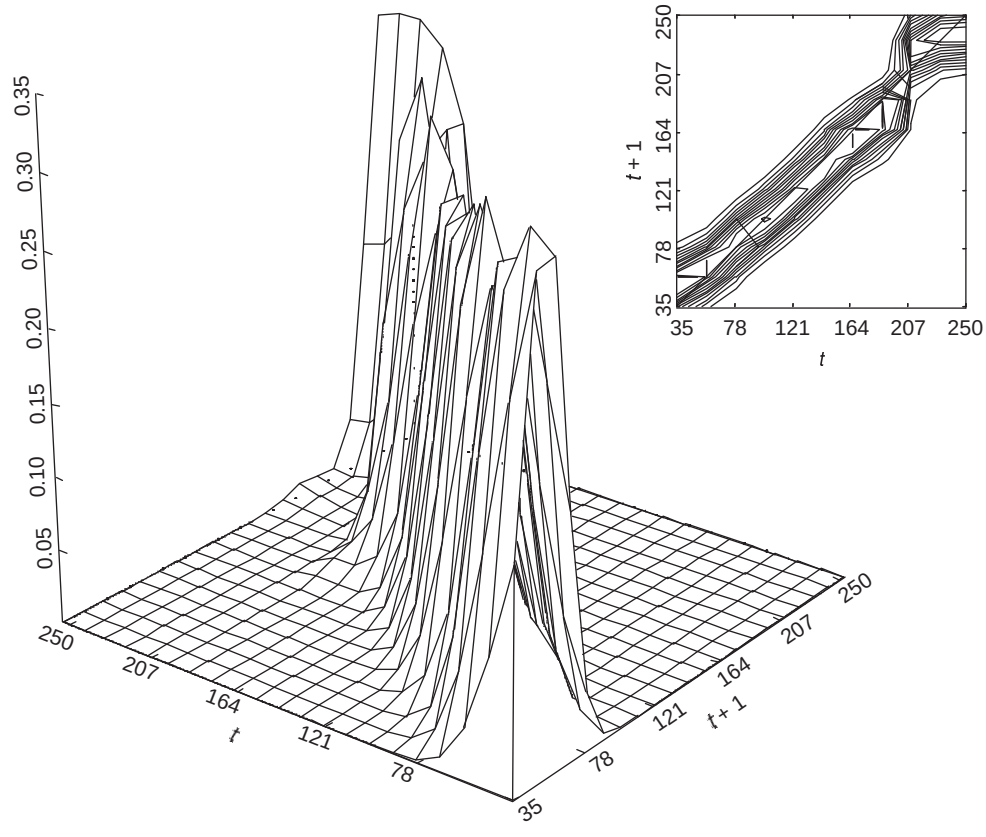


Figure 17.7 Stochastic kernel and contour plot of EU-relative regional per capita GDP distribution

of convergence around a certain per capita GDP level over time. In order to aid interpretation of the graph, Figure 17.7 also includes the related contour plot, connecting points at the same height on the three-dimensional kernel.

Figure 17.7 shows the probability mass concentrated around the main diagonal, indicating a limited mobility in the EU regional distribution of per capita GDP throughout the 23 years considered in our study. Consequently, European regions have tended on the whole to maintain their relative positions, which is consistent with the empirical evidence presented by Neven and Gouyette (1995), López-Bazo et al. (1999) and Le Gallo (2004) for a narrower geographical and time scope than ours.[15] In addition, our estimates reveal the presence of a turn above the main diagonal at the lower end of the distribution, confirming our previous findings about the relative improvement experienced by some of the least favoured areas of the EU. In contrast, the behaviour of most developed regions in Europe during the period of analysis did not contribute to a decrease in the degree of dispersion of the regional distribution of per capita GDP. Indeed, the information provided by Figure 17.7 indicates that the regions featured by a greater degree of persistence in their relative situation are indeed those located at the upper end of the distribution.[16]

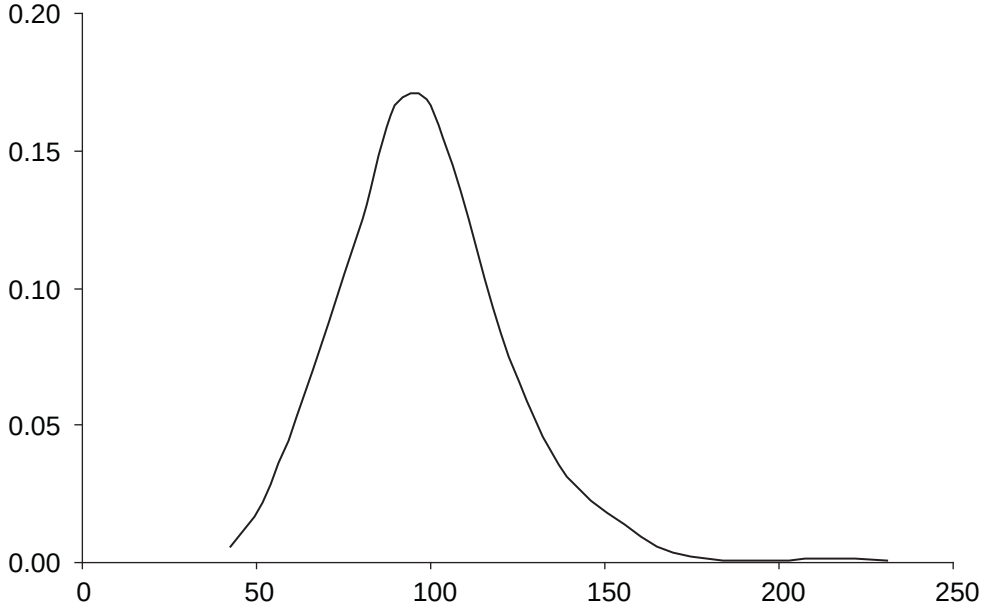


Figure 17.8 Ergodic distribution of EU-relative regional per capita GDP distribution

Finally, and in order to obtain a complete picture, we estimate the ergodic distribution by iteration of the stochastic kernel in order to reach the convergence of the process. This ergodic distribution can be interpreted as the long-run equilibrium of the regional per capita GDP distribution. Figure 17.8 shows that the estimated ergodic distribution is featured by a single mode located around the average.[17] There is thus no evidence of a potential future fragmentation into various regional clusters along different levels of economic development. This analysis also suggests that, in the long run, per capita GDP will continue at below 75 per cent of the European average for a relatively large number of regions, a factor which will need to be taken into consideration when planning the future EU regional policy.

17.6 Conclusions

In this chapter we have presented various methodological approaches to measure the regional divide. Each approach contributes to highlight different features of territorial imbalances, but ultimately only gives a partial picture of the dimension and evolution of regional disparities. There is therefore a need to combine different approaches in order to get a complete assessment.

The measures of dispersion commonly used in the literature provide a useful introduction to the quantification of the level and evolution of regional inequality. But they are based on different ethical judgements, with individual indicators aggregating the information contained in the distribution in a different way. As a consequence, diverse measures of dispersion provide different results and hence different perceptions of the dimension and evolution of regional inequality in any given geographical setting. There is therefore a need to combine several dispersion measures in order to assess the

robustness of the various indicators considered, an issue that is well known in the literature devoted to the study of personal income inequality, but that has been less salient in analyses of regional disparities.

Any reduction over time in the level of dispersion may also be compatible with the presence of polarisation processes that give rise to the formation of several internally homogenous clusters of regions. This is because the notion of polarisation is conceptually different from that of inequality and measures of inequality are inadequate to distinguish whether regions are clustered around the average of the distribution or around two or more separate poles. The evidence provided for the case of the EU demonstrates that the difference between inequality and polarisation is not simply a theoretical issue, but has practical relevance: in the period between 1980 and 2002 the EU witnessed a simultaneous reduction in the degree of regional dispersion and an increase in regional bipolarisation. This highlights the need for combining analyses of inequality and polarisation, something that has been far from common in the literature on regional disparities.

But even if a number of measures of dispersion and polarisation are combined, the resulting picture may still be partial. This problem can, to a certain extent, be addressed by the use of the non-parametric approach proposed by Quah (1993, 1996a, 1996b, 1997) to examine the evolution of the entire cross-sectional distribution. Markov chain methodologies and stochastic kernels can be applied to describe the law of motion of the distribution dynamics. This approach puts particular emphasis on two issues relevant for the analysis of regional disparities: the external shape of the distribution examined and its level of intra-distribution mobility.

One final warning is needed. As is usual in the literature, we have considered the various regions as isolated units, ignoring the spatial dimension in a strict meaning. This raises no major problems, as long as it is assumed that each regional economy evolves independently of the rest. However, this assumption is far from realistic, bearing in mind the increasing relevance of interregional flows, and the important role played by geography in the distribution of economic activity. It is therefore reasonable to suppose that neighbouring economies will enjoy similar levels of development, as has been identified by numerous studies using spatial econometric techniques (for example López-Bazo et al., 1999; Rey and Montouri, 1999) or spatial conditioning schemes in the context of the distributional approach (Quah, 1996c; Le Gallo, 2004). Hence, the spatial dimension should be seriously taken into consideration when interpreting the findings obtained by applying the various methodological approximations described in this chapter. For example, from a political perspective, the degree of inequality and polarisation needs to be considered jointly with its spatial distribution, as social unrest and political instability arising from regional disparities may be reinforced by the possible geographical concentration of poorer regions. Despite this, numerous public initiatives developed worldwide during recent decades to correct existing territorial imbalances have neglected the potential importance of spatial interaction patterns, which could affect the degree of efficiency of public intervention and our overall capacity to tackle uneven development problems

Notes

1. The study variable can be any other variable considered relevant by the researcher. Time subscripts are omitted in the rest of the chapter, unless explicitly noted.
2. In addition to sigma convergence, Barro and Sala-i-Martin (1991, 1992) introduced the notion of beta convergence, which implies the existence of an inverse relationship between the growth rate of the various

economies and their initial level of development. Since the beginning of the 1990s, a great number of studies have tested the possible presence of beta convergence in different geographical contexts by applying different approaches that include cross-sectional growth regressions, panel data techniques and time series methods.

3. This is closely related to the possible existence of convergence clubs (Baumol, 1986; Durlauf and Johnson, 1995).
4. Note that in this case there is no overlapping between the various groups, since the decomposition of the Gini index into between-group and within-group inequality is exact.
5. Some of the drawbacks of histograms include the problem of how to define the origin and length of each interval, and the possibility of improving the accuracy and efficiency of the estimates (Silverman, 1986).
6. Population size has been almost entirely ignored by the literature devoted to the non-parametric estimation of density functions using regional data (as an exception, see Ezcurra et al., 2005a).
7. The lack of complete series has obliged us to exclude from our study the countries incorporated into the EU in 2004 and 2007, the *Länder* of former East Germany, and the French Overseas Departments and Territories.
8. For example, in 2002 the Finnish region of the Åland islands had a population of 26 000 inhabitants, versus the 11 million of Île de France.
9. To check the robustness of our conclusions, we also decomposed $GE(1)$. The results were very similar. They are available from the authors upon request.
10. According to the territorial classification used by Eurostat, Denmark and Luxembourg are defined as one NUTS-2 region. It is therefore impossible to compute $GE(0)_g$ in these two cases.
11. The results for the Netherlands and Portugal should be viewed with some caution. In the case of the Netherlands, the trend followed by $GE(0)_g$ is affected by the modifications introduced in the national accounting system in the mid-1980s. The evolution of regional inequality in Portugal is influenced by the important fluctuations experienced by the GDP of the regions of Alentejo and Algarve, also during the 1980s.
12. This choice is due to the fact that, as mentioned in section 17.3, the formulation of P^{ER} is similar to that of the Gini index. The second term in expression (17.10) is the difference between two Gini indices. It is therefore reasonable to select a value of β equal to 1.
13. Estimates are based on Gaussian kernel functions, while the smoothing parameter is determined in each case following Silverman (1986), p. 48. The results obtained are robust to the kernel function used.
14. Gaussian kernel functions are used again, while the smoothing parameter is selected following Silverman (1986), p. 86.
15. There are some exceptions to this general trend. One example is the Irish region of Southern and Eastern, which showed a remarkable degree of dynamism during the 1990s. At the opposite end of the spectrum, we find several Swedish regions, such as Ovre Norrland or Mellersta Norrland, whose relative situation has deteriorated over time.
16. The estimations were repeated for different transition periods. The results in all cases were very similar to those discussed above.
17. At this point a word of warning is required: comparisons between Figure 17.8 and the density functions estimated previously should be based only on the shape of the distribution, since there is no point in comparing the level of density that appears on the vertical axis.

References

Arbia, G. (1989), *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*, Dordrecht: Kluwer Academic Publishers.

Atkinson, A.B. (1970), 'On the measurement of inequality', *Journal of Economic Theory*, **2**, 244–63.

Azzoni, C.R. (2001), 'Economic growth and income inequality in Brazil', *Annals of Regional Science*, **31**, 133–52.

Barro, R. and X. Sala-i-Martin (1991), 'Convergence across states and regions', *Brookings Papers on Economic Activity*, **1**, 107–82.

Barro, R. and X. Sala-i-Martin (1992), 'Convergence', *Journal of Political Economy*, **100**, 407–43.

Barro, R. and X. Sala-i-Martin (1995), *Economic Growth*, New York: McGraw-Hill.

Baumol, W.J. (1986), 'Productivity growth, convergence and welfare', *American Economic Review*, **76**, 1072–85.

Biewen, M. (2001), 'Bootstrap inference for inequality, mobility and poverty measurement', *Journal of Econometrics*, **108**, 317–42.

Bulli, S. (2001), 'Distribution dynamics and cross-country convergence', *Scottish Journal of Political Economy*, **48**, 226–43.

Cheshire, P.C. and G. Carbonaro (1996), 'Urban economic growth in Europe: testing theory and policy prescriptions', *Urban Studies*, **33**, 1111–28.

Cheshire, P.C. and D.G. Hay (1989), *Urban Problems in Western Europe: An Economic Analysis*, London: Unwin Hyman.
Cowell, F.A. (1995), *Measuring Inequality*, 2nd edn, LSE Handbooks in Economics, London: Prentice Hall.
Davies J.B. and A.F. Shorrocks (1989), 'Optimal grouping of income and wealth data', *Journal of Econometrics*, **42**, 97–108.
Deutsch, J. and J. Silber (1999), 'Inequality decomposition by population subgroups and the analysis of inter-distributional inequality'. in J. Silber (ed.), *Handbook of Income Inequality Measurement*, Boston: Kluwer Academic Publishers, pp. 363–404.
Durlauf, S.N. and P.A. Johnson (1995), 'Multiple regimes and cross-country growth behaviour', *Journal of Applied Econometrics*, **10**, 365–84.
Durlauf, S.N. and D. Quah (1999), 'The new empirics of economic growth', in J.B. Taylor and M. Woodford (eds), *Handbook of Macroeconomics*, Vol. 1A, Amsterdam: North-Holland, pp. 231–304.
Eckey, H.F. and M. Türk (2006), 'Convergence of EU regions: a literature report', Institut für Volkswirtschaftslhere Working Paper 80/06, Kassel Universität.
Esteban, J.M. (1994), 'La desigualdad interregional en Europa y en España: descripción y análisis', in J.M. Esteban and X. Vives (eds), *Crecimiento y convergencia regional en España y Europa*, Vol. 2, Barcelona: Instituto de Análisis Económico, pp. 13–82.
Esteban, J.M. and D. Ray (1994), 'On the measurement of polarization', *Econometrica*, **62**, 819–51.
Esteban, J.M., C. Gradín and D. Ray (2007), 'Extension of a measure of polarization with an application to the income distributions of five OECD countries', *Journal of Income Inequality*, **5**, 1–19.
European Commission (2004), *Third Report on Economic and Social Cohesion*, Brussels: European Commission.
Ezcurra, R., C. Gil, P. Pascual and M. Rapún (2005a), 'Inequality, polarisation and regional mobility in the European Union', *Urban Studies*, **42**, 1057–76.
Ezcurra, R., C. Gil and P. Pascual (2005b), 'Regional bipolarization: the case of the European Union', *International Journal of Urban and Regional Research*, **29**, 984–95.
Ezcurra, R., P. Pascual and M. Rapún (2006), 'Regional polarization in the European Union', *European Planning Studies*, **14**, 459–84.
Ezcurra, R., P. Pascual and M. Rapún (2007), 'Spatial disparities in the European Union: an analysis of regional polarization', *Annals of Regional Science*, **41**, 401–29.
Fan, C.C. and E. Casetti (1994), 'The spatial and temporal dynamics of US regional income inequality, 1950–1989', *Annals of Regional Science*, **28**, 177–96.
Foster, J. (1983), 'An axiomatic characterization of the Theil measure of income inequality', *Journal of Economic Theory*, **31**, 105–21.
Gil, C., P. Pascual and M. Rapún (2004), 'Regional economic disparities and decentralization', *Urban Studies*, **41**, 71–94.
Gradín, C. (2000), 'Polarization by sub-populations in Spain, 1973–91', *Review of Income and Wealth*, **46**, 457–74.
Johnson, P.A. (2000), 'A nonparametric analysis of income convergence across the US', *Economics Letters*, **69**, 219–23.
Le Gallo, J. (2004), 'Space–time analysis of GDP disparities among European regions: a Markov chain approach', *International Regional Science Review*, **27**, 138–63.
López-Bazo, E., E. Vayá, A. Mora and J. Suriñach (1999), 'Regional economic dynamics and convergence in the European Union', *Annals of Regional Science*, **33**, 343–70.
Maasoumi, E. (1997), 'Empirical analysis of inequality and welfare', in M. Pesaran and P. Schmidt (eds), *Handbook of Applied Econometrics*, Vol. 2, *Microeconomics*, London: Blackwell Publishers, pp. 202–45.
Magrini, S. (1999), 'The evolution of income disparities among the regions of the European Union', *Regional Science and Urban Economics*, **29**, 257–81.
Magrini, S. (2004), 'Regional (di)convergence', in V. Henderson and J. Thisse (eds), *Handbook of Urban and Regional Economics*, Vol. 4, Amsterdam: North-Holland, pp. 2741–796.
Marron, J.S. and D. Nolan (1988), 'Canonical kernels for density estimation', *Statistics and Probability Letters*, **7**, 195–9.
Mills, J.A. and S. Zandvakili (1997), 'Statistical inference via bootstrapping for measures of inequality', *Journal of Applied Econometrics*, **12**, 133–50.
Molle, W., B. Van Holst and H. Smit (1980), *Regional Disparity and Economic Development in the European Community*, Farnborough: Saxon House.
Neven, D. and C. Gouyette (1995), 'Regional convergence in the European Community', *Journal of Common Market Studies*, **33**, 47–65.
Ottaviano, G. and D. Puga (1998), 'Agglomeration in the global economy: a survey of the new economic geography', *World Economy*, **21**, 707–31.
Overman, H. and D. Puga (2002), 'Unemployment clusters across Europe's regions and countries', *Economic Policy*, **34**, 115–47.

Pekkala, S. (2000), 'Aggregate economic fluctuations and regional convergence: the Finnish case 1988–95', *Applied Economics*, **32**, 211–19.
Petrakos, G., A. Rodríguez-Pose and A. Rovolis (2005), 'Growth, integration and regional inequality in Europe', *Environment and Planning A*, **37**, 1837–55.
Puga, D. (2002), 'European regional policies in light of recent location theories', *Journal of Economic Geography*, **2**, 373–406.
Quah, D. (1993), 'Empirical cross-section dynamics in economic growth', *European Economic Review*, **37**, 426–34.
Quah, D. (1996a), 'Twin peaks: growth and convergence in models of distribution dynamics', *Economic Journal*, **106**, 1045–55.
Quah, D. (1996b), 'Empirics for economic growth and convergence', *European Economic Review*, **40**, 1353–75.
Quah, D. (1996c), 'Regional convergence clusters across Europe', *European Economic Review*, **40**, 951–8.
Quah, D. (1997), 'Empirics for growth and distribution: stratification, polarization and convergence clubs', *Journal of Economic Growth*, **2**, 27–59.
Rey, S. (2001), 'Spatial analysis of regional income inequality', mimeo, San Diego State University, http://ideas.repec.org/p/wpa/wuwpur/0110002.html.
Rey, S. and B.D. Montouri (1999), 'US regional income convergence: a spatial econometric perspective', *Regional Studies*, **33**, 143–56.
Rodríguez-Pose, A. (1998), *Dynamics of Regional Growth in Europe: Social and Political Factors*, Oxford: Clarendon Press.
Rodríguez-Pose, A. (1999), 'Convergence or divergence? Types of regional responses to socio-economic change in Western Europe', *Tijdschrift voor Economische in Sociale Geografie*, **90**, 363–78.
Sen, A. (1973), *On Economic Inequality*, Oxford: Oxford University Press.
Shorrocks, A.F. (1980), 'The class of additively decomposable inequality measures', *Econometrica*, **48**, 613–25.
Shorrocks, A.F. (1984), 'Inequality decomposition by population subgroups', *Econometrica*, **52**, 1369–85.
Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability 26, London: Chapman and Hall.
Terrasi, M. (1999), 'Convergence and divergence across Italian regions', *Annals of Regional Science*, **33**, 491–510.
Theil, H. (1967), *Economics and Information Theory*, Amsterdam: North-Holland.
Williamson, J.G. (1965), 'Regional inequality and the process of national development: a description of patterns', *Economic Development and Cultural Change*, **13**, 3–45.
Wolfson, M. (1994), 'When inequalities diverge', *American Economic Review*, **84**, 353–8.
Zhang, X. and R. Kanbur (2001), 'What difference do polarization measures make?', *Journal of Development Studies*, **37**, 85–98.

18 Measuring regional endogenous growth

Robert J. Stimson, Alistair Robson and Tung-Kai Shyy

18.1 Introduction

The importance of endogenous factors in regional economic development and growth has long been recognized and is largely the basis of the so-called 'new growth theory' that has been popularized since the 1980s. Theories of regional endogenous growth place emphasis not only on regional resource endowments and human capital – which have always been viewed as important factors affecting the economic development of regions – but also on technology, entrepreneurship and institutional factors, including the role of leadership. However, there is neither a standard definition of endogenous growth, nor a specification of an operational model including those factors that determine spatial variations in regional endogenous growth performance. In fact, there is no universally accepted measure of endogenous growth.[1]

This chapter first provides an overview discussion of endogenous growth factors. It then proposes a measure of regional endogenous change which is readily calculable from secondary analysis of regional employment data available in the national census. The regional or differential/regional shift component derived from shift-share analysis of employment change over time is proposed as a viable proxy measure as a dependent variable in an endogenous growth model. A series of independent variables which also may be derived from census data are specified in the model as factors likely to explain spatial variability in regional performance on that dependent variable. Those variables are taken as reflecting the types of factors that are proposed in the regional economic development literature as potentially influencing endogenous growth. The results derived from the application of the model across non-metropolitan regions in the state of Queensland, Australia are presented. The chapter concludes with some thoughts on the emergence of a new paradigm for regional economic development analysis and planning.

18.2 The shift from an exogenous to an endogenous focus

The evolution of theories on regional economic growth and development has seen a shift in the emphasis from exogenous to endogenous factors. Traditional regional economic development approaches were embedded in neoclassical economic exogenous growth theory, based largely on the Solow (1956, 2000) model. These have been replaced by a suite of models and arguments that are commonly known as the new growth theory where the focus is directed towards endogenous factors and processes (see, for example, Johansson et al., 2001). Some of the key contributions to this shift in emphasis towards endogenous processes as drivers in regional growth and development are briefly discussed in what follows.

Traditionally, most researchers have understood endogenous growth as attempts to endogenise technology (see, for example, Romer, 1990) and human capital (see, for example, Lucas, 1985, 1988). This leads to the possibility of increasing returns to scale and divergence. In this section we extend this definition to incorporate some other

endogenous-related concepts, such as product cycle theory (see Norton and Rees, 1979), innovative milieux (see Saxenian, 1994), learning regions (see Simmie, 1997), regional systems of innovation, new industrial spaces (see Scott, 1988), new economic geography (see Krugman, 1991), trust (see Fukuyama, 1995) and competitive advantage (see Porter, 1990). These have all become increasingly incorporated under the 'endogenous regional growth theory' umbrella.

Emergence of the 'new growth theory'

In general economic theory, 'endogenous growth theory' (also referred to as 'new growth theory') has been intensively studied since Romer's (1986) seminal paper which founded the theory. It was aimed at explaining why there is a divergence in incomes between countries (that is, the origin of growth), and emerged during the second major period of research into economic growth theory during the 1980s as a response to criticism of neoclassical growth theories. The other major growth theory – exogenous growth theory – was founded much earlier by the likes of Solow (1956), Cass (1965) and Koopmans (1965) in the first major period of interest in economic growth theory. Both foundation theories have been modified over time. Notable updates of the exogenous growth theory have been made by Parente and Prescott (1994) and Ngai (2004). Without delving too heavily into the theory, which would take a vast amount of space, it has been claimed by numerous critics of endogenous growth theory (such as Parente, 2001) that endogenous growth theories have failed to explain non-convergence between countries. This finding could have implications for regional economic growth theory, which appears to increasingly rely on endogenous growth theory.

As with many other schools of economic thought there are numerous variants of endogenous growth theory, such as Romer's (1986) competitive equilibrium model, and Lucas's (1985) 'new classical' version. One of the common key outcomes of endogenous growth theory is that policy measures can impact on the long-run growth rate of the economy. This is often achieved through higher savings and/or investment levels, new technology and human capital (that is, endogenise technology and human capital) which increase returns to scale and hence divergence in economic performance. In exogenous growth models, higher savings and investment do increase economic growth, but only in the short term. For economic geographers, the ability to affect long-run economic growth via policy measures is attractive, and may partially explain the popularity of endogenous growth theory compared with exogenous growth theory.

In the evolution of regional economic theory, in the 1970s Rees (1979) had proposed that technology was a prime driver in regional economic development, and since then over the ensuing two to three decades the regional science literature has shown how technology is directly related to traditional concepts of agglomeration economies in regional economic development, and to new or repackaged older concepts of entrepreneurship, institutions, and leadership. From that time theorists such as Romer (1986, 1990), Lucas (1985), Barro (1990), Rebelo (1991), Grossman and Helpman (1991) and Arthur (1994) sought to explain technical progress as it generates economic development as an endogenous effect rather than accepting the neoclassical view of long-term growth being due to exogenous factors. Thomas (1975) and later Erickson (1994), among others, showed how technological change is related to the competitiveness of regions. And Norton and Rees (1979) and Erickson and Leinbach (1979) showed how the product

cycle, when incorporated into a spatial setting, may impact differentially on regions through three stages, namely:

- an innovation stage;
- a growth stage;
- a standardization stage.

In this transition, production can shift from the original high-cost home region to a lower-cost location. This has often been offshore, which has been hastened through the evolution of the internationalization of the production process.[2] Thus some regions are the innovators, while others become the branch plants or recipients of the innovation production, and these might even then become innovators via endogenous growth. Markusen (1985) extended the product cycle theory of regional economic development by articulating how profit cycles and oligopoly in various types of industrial organization and corporate development can magnify regional economic development differentials.

The concept of the 'innovative milieu' was formulated to explain the 'how, when and why' of new technology generation. That notion linked back to the importance of agglomeration economies and localization economies that may lead to the development of new industrial spaces (Scott, 1988; Porter, 1990; Krugman, 1991). This relates to the possibility of increasing returns to scale and divergence discussed by Romer and Lucas (and referred to previously) in endogenous growth theory.

Some theorists, such as Fukuyama (1995), have suggested that not just economic but also value and cultural factors – including social capital and trust – are important in the rise of technology agglomerations as seen in the Silicon Valley phenomenon. Collaboration among small and medium-sized enterprises through networks and alliances and links with universities forged a powerful research and development (R&D) and entrepreneurial business climate. Nonetheless, Castells and Hall (1994), in discussing innovative industrial milieu, note the following: 'despite all this activity . . . most of the world's actual high-technology production and innovation still comes from areas that are not usually heralded as innovative milieus . . . the great metropolitan areas of the industrial world' (p. 11). However, as Rees (2001) points out, technology-based theories of regional economic development need to incorporate the role of entrepreneurship and leadership, particularly as factors in the endogenous growth of regions, and it is the 'link between the role of technology change and leadership that can lead to the growth of new industrial regions and to the regeneration of older ones' (p. 107).

Thus, in the new growth theory models, allowance is made for both agglomeration effects (economies of scale and externalities), and market imperfections, with the price mechanism not necessarily generating an optimal outcome through efficient allocation of resources in the long run. Also, the processes of capital accumulation and free trade do not necessarily lead to convergence between regions, with positive agglomeration effects concentrating activity in one or a few regions through self-enforcing effects that attract new investment. Most importantly, the new growth theory allows for both concentration and divergence.

Factors such as those referred to above are being seen as fundamental processes affecting regional economic growth and development, arising from the resource endowments and the

knowledge base of a region, and being enhanced through entrepreneurship, innovation, the adoption of new technologies, leadership, institutional capacity and capability, and learning.

As part of this evolution in regional growth and development theory, there has been also been a shift from concerns about developing a regional comparative advantage to developing a regional competitive advantage. There has also been a shift in regional development planning strategy from master planning and structural planning to strategic planning paradigms – a new way of conceptualizing regional economic growth and development has emerged within the 'new growth theory', thus extending it from the pure Lucas and Romer concepts of endogenous growth discussed previously.

Stimson et al. (2005) have proposed a new model framework which explicitly incorporates a wide set of endogenous factors as intervening variables that are hypothesized to have a catalytic impact on the regional endogenous economic growth and development process. They propose an operational model may be represented as:

$$RED = f[RE, M \text{ ...mediated by ... } L, I, E)]$$

Here the outcome of the regional economic development process (*RED*) is the degree to which a region has achieved competitive performance, displays entrepreneurship, and has achieved sustainable development. Those outcome states are defined as the dependent variable(s) in the model. An outcome state is conceptualized as being dependent on a set of quasi-independent variables relating to a city or region's resource endowments (*RE*) and its 'fit' with market conditions (*M*), that being mediated through the interaction between sets of intervening variables that encompass factors defined as leadership (*L*) and institutions (*I*) which may interact to facilitate or suppress entrepreneurship (*E*). This new model framework incorporates both direct and indirect effects in the interactions between *RE*, *M* (the quasi-independent variables) and *L*, *I* and *E* (the intervening or mediating variables). Also, the interactions between the intervening or mediating variables *L*, *I* and *E* may be both direct and indirect.

It is thus evident that there have emerged a number of key themes as to what constitutes regional economic growth and development and regional competitiveness. Not surprisingly there are differences of views among regional economic development scholars, and some of those differences relate to the relative focus given to the roles of exogenous forces on the one hand and the roles of endogenous processes and factors on the other hand. But there does now seem to be an almost universal realization of what Garlick et al. (2006) refer to as the institutional embeddedness of endogenous processes and factors in regional development.

Of course exogenous factors do remain important to a region's economic performance and how it develops over time; but increasing importance is being placed on endogenous forces as determinants of a region's competitiveness. However, regional economic development policy initiatives now tend to be more oriented towards measures that enhance local capacity and capability for a city or region to develop and cope with rapid change in an increasingly competitive global environment. While endogenous growth theory makes mention of leadership, entrepreneurship and institutional factors, little systematic analysis has occurred to thoroughly conceptualize or, even more, measure their roles as endogenous factors in the development process.

A welcome development for regional economic development planners

These evolutionary developments in regional economic growth and development theory discussed above are welcome for regional economic development analysts and planners because, among other things, they explicitly introduce a spatial dimension into economic growth theory, a dimension which was ignored in neoclassical economic development theory. That evolution is particularly important as the role of regions in national economies has changed significantly since the 1970s. This has largely been a result of globalization and structural change and adjustment. Understanding those processes of change is crucial for analysing and understanding differentials in the patterns of regional economic performance and for formulating and implementing regional economic development planning strategy.

As discussed by Stimson et al. (2006), the challenge facing economic development planners in contemporary times is how to formulate economic policy that will respond to global dynamics, and sometimes (or often) a national vacuum in macro policy towards regions in many countries.

At one time regions were protected from outside competition, and to some extent their economies could be manipulated by national governments. But that is no longer the case as the economic rationalism[3] pursued by many national governments left many regions to fend for themselves. Many regions still continue to look to higher levels of government for support and resources to provide economic direction and investment to stimulate economic development. However, many regions fail to understand that globalization has left such governments largely devoid of powers to apply economic and policy mechanisms to enhance the competitiveness of regional economies. As discussed by Stimson et al. (2006), today it is increasingly up to regions to develop and use their own devices to compete internationally in order to survive. That is, a reliance on endogenous processes is typically espoused in regional economic development policy. To do so regions need first to understand what the factors are that set the dynamics of the emerging new economic age of the twenty-first century.

Some findings from empirical analyses of endogenous factors associated with regional growth

Differentials in the pattern of regional economic performance discussed here may be attributed to many factors, both exogenous and endogenous to a region. The increasing focus on endogenous factors tends to canvass issues such as regional economic diversity, population size and agglomeration, levels of human capital, and income. Numerous cross-sectional techniques, as well as portfolio theoretic approaches, have been used by researchers to investigate such relationships.

Much of the recent empirical analysis and modelling of regional performance has tended to pay specific attention to investigating the influence of endogenous factors relating to aspects of regional industrial structure and human capital.

Regarding the effect of industrial structure on regional stability and growth, one argument is that industrial diversity and a trend towards diversification of industry sector employment enhances opportunities for growth and development (see, for example, Henderson et al., 1995; Gordon and McCann, 2000), although some researchers claim that there is scant empirical evidence for that proposition (Kaufman, 1993; Lande, 1994; Productivity Commission, 1998). There is also debate on the influence of urban scale and agglomeration (Taylor et al., 2002) on regional performance. A study by Duranton and

Puga (2000) reviewing national urban systems suggested that larger cities are more diversified, that individual city-size rankings and individual city specialization tend to be stable over time, and that specialised and diversified cities co-exist across a national urban system.

In Australia, empirical analysis by the Bureau of Transport and Regional Economics (2004a) shows that there has emerged a more diverse industry structure outside the major cities over the decade 1991 to 2001. Approximately 20 per cent of the variation in employment rates across the nation's labour market regions may be explained by the industry structure of regional employment at the beginning of the decade 1991–2001. Potentially, other important influences on regional economic growth might include endogenous factors such as amenities, remoteness, investment leadership and a region's resource base and skills base. For example, Garnaut et al. (2001) have found that metropolitan and coastal regions experience stronger population and employment growth over time compared to inland and remote areas; and Bradley and Gans (1998) have found that labour force growth is correlated positively to levels of industrial diversity and negatively to the initial size of a city. Lawson and Dwyer (2002) have found that regions with high levels of employment in accommodation, cafes and restaurants, or with high industrial diversity at the start of a period, have higher employment growth.[4] A higher level of industrial diversity also has a positive effect on levels of regional growth. In addition they found that regions experiencing high rates of structural change were more likely to have recorded employment growth. But less significant is the likelihood that employment growth is associated with a coastal location.

In addition, there has also been much discussion of the important effects of human capital skills and income in explaining differential levels of regional economic performance (see, for example, Hanushek and Kimko, 2000; Goetz and Rapasingla, 2001). In Australia, empirical analysis by Draca et al. (2003) shows how levels of education, skills and qualifications explain between 10 per cent and 20 per cent of the variation between the states in gross state product in Victoria, Queensland and South Australia. Research by Lawson and Dwyer (2002), Harrison (1997), Norris and Wooden (1996), Garnett and Lewis (2000) and Stimson et al. (2004) points to the effect of geographically uneven patterns in the level of skills and qualifications across Australia's regions being related to differential levels of population growth (and decline) and issues of employment across industry sectors. However, the Bureau of Transport and Regional Economics (2004b) research suggests the links between education, labour quality and productivity, and regional growth are complex, but that the level of human capital does seem to be related to well-being and regional productivity.

18.3 An approach to measuring and modelling regional endogenous growth

Despite the evolving focus on endogenous factors in the regional economic growth and development literature, it is somewhat surprising that there is no standard definition of endogenous regional economic growth (or decline) in terms of the specification of an agreed variable which measures it. Thus, a key issue is: What is an appropriate proxy measure of a region's endogenous growth?

Furthermore, it is also surprising that there is a lack of operational models to measure the effect of factors such as those discussed above on explaining spatial variations in regional endogenous growth performance in a nation or state.

A definition of regional employment growth: a proxy measure for the dependent variable
In general economic analysis, most studies derive and measure a variable for endogenous growth by using ordinary least squares or, more recently, panel data analysis. However, such data are not usually readily available at the disaggregated regional level across a nation. As such, many economic geographers and regional economists have turned to other techniques. One such approach used to measure regional endogenous economic growth and development has been proposed by Stimson et al. (2005). It takes a proxy measure such as:

- the aggregated (across all industry sectors) regional differential shift component value in a shift-share analysis, a common technique in analysing regional differential performance used by economic geographers and regional economists; or
- an employment scale weighted location quotient change over time, standardized by the size of the region's labour force.

Such a measure can then be used as the dependent variable in a model of regional endogenous economic growth and development.

In the model application discussed in this chapter, we have attempted to combine a range of endogenous factors that reflect economic growth (measured by employment, given data constraints on measuring production at the regional level) rather than one measure, such as technology or level of savings. We do this by using shift-share analysis as proposed by Stimson et al. (2005). One reason for adopting this approach is that secondary data tend to be readily available in most countries to perform a shift-share analysis, and typically that may be achieved using census data for industry employment in regions. The regional shift component is a reasonable surrogate measure of the degree to which employment growth or decline in a region is due to endogenous or 'within-region' processes and to exogenous factors, such as national shift and the industry-mix shift effects. Indeed, that is what the regional shift component is purported to measure. While the authors recognize that using the differential/regional shift component from a shift-share analysis is not an ideal method of measuring regional endogenous regional economic growth (or decline), it is nonetheless considered to be the most optimal given the lack of data at the regional level to operationalize other measurement approaches.

Unfortunately it has not been possible to use the full new model framework proposed by Stimson et al. (2005) because of the non-availability of regional-level data for the intervening variables relating to leadership, institutional and entrepreneurship factors. Thus a more restricted model has been operationalized as explained in the sections that follow.

A model application: regional endogenous growth in non-metropolitan Queensland, Australia
The above approach has been applied in an exploratory study modelling regional endogenous employment growth in Queensland, Australia by Stimson et al. (2004). It uses the Haynes and Dinc (1997) method to derive the differential/regional shift component (designated REG_SHIF) in a shift-share analysis of employment change – additive across all industry sectors – in non-metropolitan local government areas (LGAs) between the 1991 and 2001 censuses. Because of the large variation in the size of LGAs across rural and

regional Queensland – from the most populous LGA, Gold Coast City, with a population of over 418 000 in 2001, to several LGAs in the inland western parts of the state with populations of under 1000 in 2001 – it was necessary to standardize the REG_SHIF variable to account for the size of the LGA labour force in 1991. This is used as the dependent variable in the model which seeks to identify factors that explain spatial variations in regional endogenous employment growth performance across that state's non-metropolitan LGAs.

The choice of independent variables was guided by the findings in the literature on regional economic growth and development, such as in the literature referred to in the previous section. In their study, Stimson et al. (2004) were confined to using variables that could be derived from the census data for 1991 and 2001. A total of 27 independent variables were selected for use in the model. They are listed in Table 18.1, and the reason for their choice is discussed below.

Industrial structure and size effects A series of variables were compiled to enable us to test the impact on regional endogenous growth of both industrial diversity and size. This was addressed first by compiling an industry specialization index for 1991 (SPEC_91) and for 2001, and then calculating a measure of change in the specialization index over the decade (SPEC_CH). For this purpose employment data in LGAS in 17 broad industry sectors were used. Second, a structural change index was computed for 1991 to 2001 (SCI_91-01), and a measure of change in the structural change index for the period from 1991–96 to 1996–2001 was compiled (SCI_CH).

The log of the population of LGAs in 1991 (L_POP_91) was used as a measure of size of a region at the beginning of the decade, and the percentage point growth (or decline) in population size over the decade 1991–2001 (POP_CH) was used as a dynamic measure of local market size. In order to address the effect of proximity to the metro-region, a dummy variable (D_METRO) was used to indicate whether an LGA was adjacent to the Brisbane Statistical Division.

To investigate further the nature of the industry structure of LGAs and the effect of industry specialization on regional endogenous growth, a location quotient for three key industry sectors in 1991 was calculated for each LGA: namely, for employment in manufacturing (LQ_MAN_91), for property and business services (LQPBS_91) and for personal and other services (LQPER_91). In addition, variables measuring the change in the location quotient of employment in these three industry sectors over the decade 1991–2001 were derived (LQMAN_CH, LQPBS_CH, and LQPER_CH).

Labour force utilization To investigate the effect on endogenous growth of labour force utilisation, a simple measure of the unemployment rate in 1991 (UNEMP_91) was used, along with a measure of change in the unemployment rate over the decade 1991–2001 (UNEMP_CH).

Human capital and income It is often proposed that regional economic growth is enhanced by the existence of skilled workers, the availability of employment opportunities, opportunities for a wide range of skills, and the existence of higher-income jobs. A number of variables were incorporated in the model to assess those effects. First, a variable was derived to measure income levels by taking the log of average annual income[5] for

Table 18.1 Definition of variables used in the Queensland study model

Variable	Definition
REG_SHIF	Regional shift (from 1991 to 2001)/Labour Force (1991) (i.e. employment + unemployed)
SPEC_91	Specialization index for 1991 across 17 industry sectors
SPEC_CH	Change in specialization index from 1991 to 2001
SCI_91_01	Structural change index for 1991 to 2001
SCI_CH	Change in the structure change index from 1991−96 to 1996–2001
L_INC_01	Log(average annual income for 2001)
UNEMP_91	Unemployment rate in 1991 (for all persons)
UNEMP_CH	Unemployment rate change from 1991 to 2001 (for all persons)
L_POP_91	Log(Population for all persons in 1991)
POP_CH	Percentage point change in population from 1991 to 2001
LQMAN_91	Location quotient for the manufacturing industry in 1991
LQMAN_CH	Change in the location quotient for the manufacturing industry from 1991 to 2001
LQPBS_91	Location quotient for the property and business services industry in 1991
LQPBS_CH	Change in the location quotient for the property and business services industry from 1991 to 2001
LQPER_91	Location quotient for the personal and other services industry in 1991
LQPER_CH	Change in the location quotient for the personal and other services industry from 1991 to 2001
UNIQUALS_91	Proportion of the population with a Bachelors degree or higher in 1991
UNIQUALS_CH	Change in the proportion of the population with a Bachelors degree or higher from 1991 to 2001
TECHQUALS_91	Proportion of the population with a technical qualification in 1991
TECHQUALS_CH	Change in the proportion of the population with a technical qualification from 1991 to 2001
ROUTW_91	Proportion of total occupations (all persons) as routine production workers for 1991
INPERS_91	Proportion of total occupations (all persons) as in-person service workers for 1991
SYMBA_91	Proportion of total occupations (all persons) as symbolic analysts for 1991
ROUTW_CH	Change in proportion of total occupations (all persons) as routine production workers from 1991 to 2001
INPERS_CH	Change in proportion of total occupations (all persons) as in-person service workers from 1991 to 2001
SYMBA_CH	Change in proportion of total occupations (all persons) as symbolic analysts from 1991 to 2001
D_COAST	Border is adjacent to coastline (dummy variable 0 = NO, 1 = YES)
D_METRO	Adjacent to the state capital statistical district (SD) (dummy variable 0 = NO, 1 = YES)

Source: Stimson et al. (2004).

an LGA at 2001 (L_INC_01). Then a series of measures of the level of human capital were derived, namely the proportion of an LGA's population with a bachelors or higher degree qualification in 1991 (UNIQUALS_91) and the proportion with a technical qualification (TECHQUALS_91). To measure the effect of shifts in those levels of human capital, two variables were created on the change from 1991 to 2001 in the incidence of those qualifications (UNIQUALS_CH and TECHQUALS_CH).

Occupational shifts It has been argued by Reich (1991) and others that the evolution of the knowledge-based economy and of information-intensive activities has led to a restructuring of occupations vis-à-vis skills and functions. Thus, it was decided to reorganise the 1991 census data on the occupational structure of LGAs into three broad groupings that represent Reich's (1991) symbolic analysis, in-person service workers, and routine production workers (SYMBA_91, INPE_91, and ROUTW_91). A measure of change over the decade 1991–2001 in employment in those categories was also calculated (SYMBA_CH, INPERS_CH, and ROUTW_CH).

Coastal and inland effects Some of the research in Australia inquiring into regional economic performance has investigated the spatial patterns of variations in the context of geographic variables such as remoteness, and coastal and inland environments and locations. To at least incorporate the potential effect of a coastal location on regional employment endogenous growth (or decline), a dummy variable was included to indicate whether or not an LGA is located adjacent to the coast of Queensland (D_COAST).

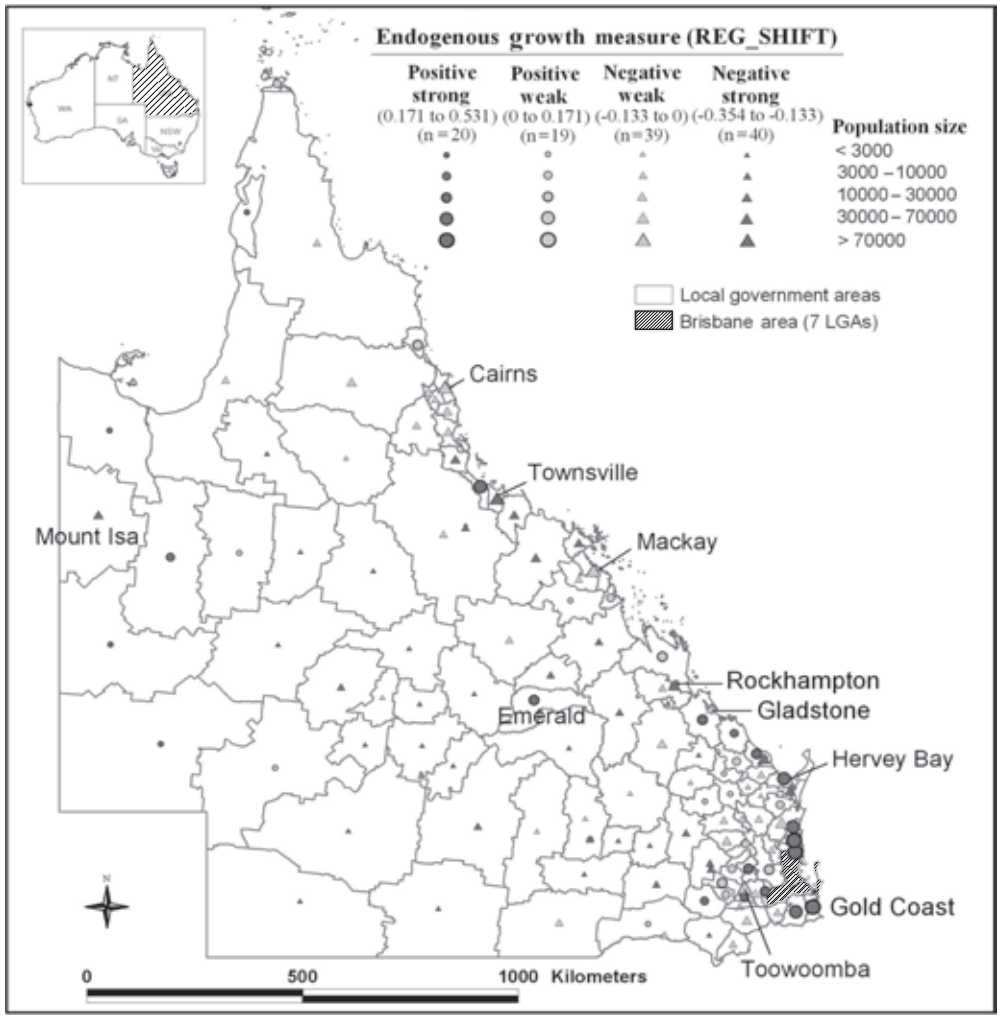
Model estimation and evaluation

Having determined the variables to be included in the general model, Stimson, et al. (2004) first used a conventional ordinary least squares (OLS) technique to estimate the model of regional endogenous growth across Queensland's rural and regional LGAs. Then a stepwise approach was employed to determine a specific model. Each step involved withdrawing one independent variable from the model. The variable deleted had the highest probability that its absolute p-value[6] was greater than 0.05.[7] The model was then regressed and new estimates of the model were obtained. This process occurred until a variable was identified with the highest probability that its absolute p-value was less than 0.05.[8] Here only the results of the specific (stepwise) model are discussed.

One key feature for a reasonable OLS regression is non-constant variance in the residuals. A test was run to evaluate this on the estimated variable. Tests were also run to determine the existence of spatial autocorrelation.

Patterns of regional endogenous employment growth/decline performance

The spatial patterns of variation across Queensland's non-metropolitan LGAs in their performance on the REG_SHIF dependent variable used in the model as the proxy measure of regional endogenous growth are presented in Figure 18.1. Rather than using a choropleth map, Stimson et al. (2004) used symbols (placed at the centroid of each LGA) to represent the magnitude of an LGA's positive or negative score on the regional employment shift (REG_SHIF) dependent variable. Those symbols are graduated in scale in order to represent the size category for the population of an LGA. A circle symbol represents a positive score on the regional employment shift variable, indicating that an LGA



Source: Stimson et al. (2004).

Figure 18.1 Spatial pattern of LGA performance on the regional endogenous growth variable

has experienced employment growth over the decade on the regional shift component derived from the shift-share analysis, standardized by size of the labour force of the LGA in 1991. A triangle symbol represents a negative score on the REG_SHIF variable, indicating loss of jobs over the decade 1991–2001 due to endogenous processes and factors. The map also uses black and white circles (positive) and triangles (negative) to indicate the magnitude of the endogenous growth or decline effect. Some of the larger regional cities and towns are identified.

Overall, 39 of Queensland's non-metropolitan LGAs are shown to have experienced endogenous growth in employment over the decade 1991–2001, while a total of 79 LGAs

display a decline in employment due to endogenous processes and factors. Figure 18.1 clearly shows that endogenous employment growth is most evident in some of the larger coastal LGAs in the south-eastern corner of the state adjacent to and inland from the Brisbane metro-area, and as well across a few of the inland small-population rural LGAs and some of the inland regional centres.

Figure 18.1 also shows that negative scores indicating employment decline due to endogenous processes and factors are widely apparent across much of western rural and regional Queensland, and as well in many of the LGAs (particularly those with smaller populations) along the coast and in near coastal locations.

The model results

The general model (including all variables) showed that the variable with the greatest influence on the regional shift variable was found to be population change (POP_CH), which has a very significant effect in a positive direction. Average income (L_INC_01) also has a strong positive effect on endogenous growth. The change in unemployment from 1991 to 2001 (UNEMP_CH) has had a strong negative effect. Both the proportion of the population with university (UNIQUALS_91) and with technical qualifications (TECHQUALS_91) in 1991 have a strong negative influence. The specialization index in 1991 (SPEC_91), and the change in that index from 1991 to 2001 (SPEC_CH), both have a strong positive influence on endogenous growth.

The results of the stepwise process are presented in Table 18.2. This estimates the specific model. The explanatory power of this specific model is very similar to the general model (adjusted R-squared value 0.89), but the benefit is the higher degrees of freedom in the model. There are 12 significant independent explanatory variables which affected regional endogenous growth (REG_SHIF). The new additional variables in Table 18.2, as compared to the general model, are: population size in 1991 (L_POP_91); the location quotient for employment in personal and other services in 1991 (LQPERS_91); the change in the proportion of the population with university qualifications from 1991 to 2001 (UNIQUALS_CH); and the change from 1991 to 2001 in the proportion of the labour force who were routine workers and in-person services workers (ROUTW_CH). Most of these explanatory variables have a significant positive relationship with the dependent variable. However, the change in unemployment from 1991 to 2001 (UNEMP_CH), and the proportion of the population with a university degree in 1991 (UNIQUALS_91) and with technical qualifications (TECHQUALS_91) have a significant negative relationship.

These results need to be qualified by examining the co-linearity between the independent variables as shown by an intercorrelation matrix. The highest co-linearity is between SPEC_91 and L_POP_91 (-0.65), L_POP_91 and TECHQUALS_91 (0.57), and UNIQUALS_91 and TECHQUALS_91 (0.57). This co-linearity could affect the results of the analysis.

The data used in the modelling are geographically related, and because of that spatial autocorrelation could emerge in the estimated model. To examine for spatial autocorrelation, Stimson et al. (2004) conducted a number of tests, including the Moran *I* statistic, Lagrange multiplier (LM) error, and LM lag tests. The Moran *I* test is not significant. Therefore the null hypothesis of spatial independence cannot be rejected. Although the Moran *I* statistic is a useful and easy method to apply, as well as being consistently the optimal method of determining spatial independence, it does not offer an explanation of

Table 18.2 OLS specific model results

	Estimate		t value		
	Sign	Value	Sign	Value	Pr(<|t|) value
(Intercept)	−	3.43	−	9.39	1.37e–15
SPEC_91	+	0.27	+	2.41	0.02
SPEC_CH	+	0.45	+	2.12	0.03
L_INC_01	+	0.73	+	8.85	2.21e–14
UNEMP_CH	−	1.10	−	3.88	0.00
L_POP_91	+	0.04	+	2.66	0.01
POP_CH	+	0.80	+	22.19	<2e-16
LQPER_CH	+	0.01	+	2.26	0.02
UNIQUALS_91	−	2.58	−	3.19	0.01
UNIQUALS_CH	+	1.88	+	2.39	0.02
TECHQUALS_91	−	1.635	−	4.64	1.00e-05
ROUTW_CH	+	0.79	+	3.66	0.00
INPERS_CH	+	1.17	+	4.07	8.90e-05

Notes: Residual standard error: 0.06 on 105 degrees of freedom; Multiple R-Squared: 0.89, Adjusted R-squared: 0.88; F-statistic: 77.52 on 12 and 105 DF, p-value: <2e-16.

Source: Stimson et al. (2004).

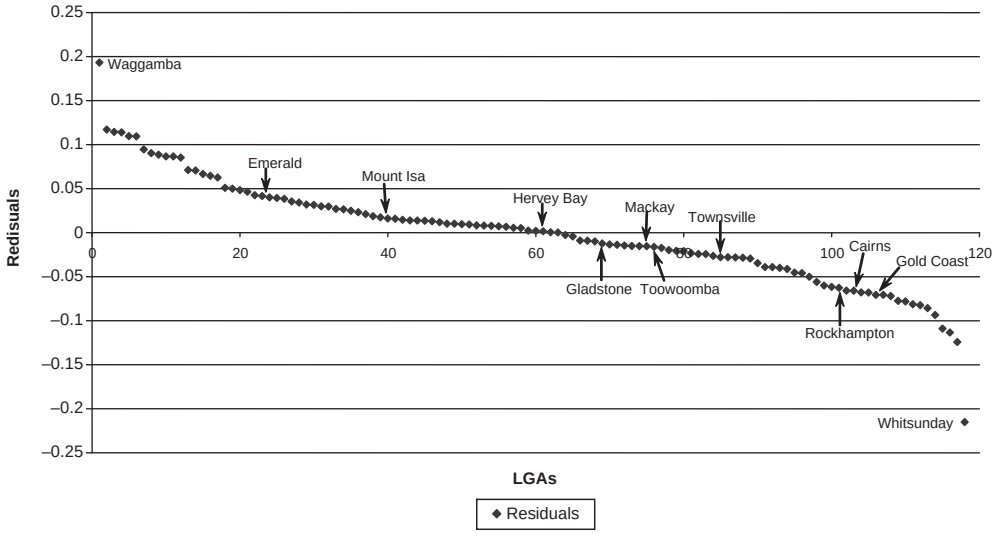
whether this autocorrelation occurs from a true spatial process or an error process (Anselin et al., 1996). To address these problems, the LM error and LM lag tests were introduced in this test, allowing an indication of which process is best represented in the general model. However, given that the Moran *I* test statistic indicated no spatial autocorrelation, those other tests were not utilized.

A feature for a reasonable OLS regression is non-constant variance in the residuals. This was not found to be a problem with the estimated model in the specific model. The line of the fit of the residuals to the regression of the specific model showed no visual relationship between the residuals and the fitted regression line.

Patterns of residuals

Figure 18.2 plots the deviation of the residuals for LGAs from the line of best fit derived from the specific model. It is evident there are two outlier LGAs: Whitsunday (negative) located on the central coast barrier reef, and Waggamba located in inland southern Queensland. The positions of the major regional centres are indicated on this plot. It is evident that the large majority of residual scores for the LGAs fall within the range of +0.05 to −0.05.

The spatial patterns of the residuals are mapped in Figure 18.3. There are 79 LGAs which fall within the +0.05 to -0.05 range of residuals. There are 45 LGAs for which the model estimates are positive or above the regression line, and these are largely found to be located around the Brisbane metro-area in the south-east of Queensland, while some such LGAs are also small places located across the inland areas of the state. The 34 LGAs for which the model estimates are negative or below the regression line within this range are typically larger coastal centres or small rural places on the coast, and as well some of them are found in the inland south-eastern parts of the state.



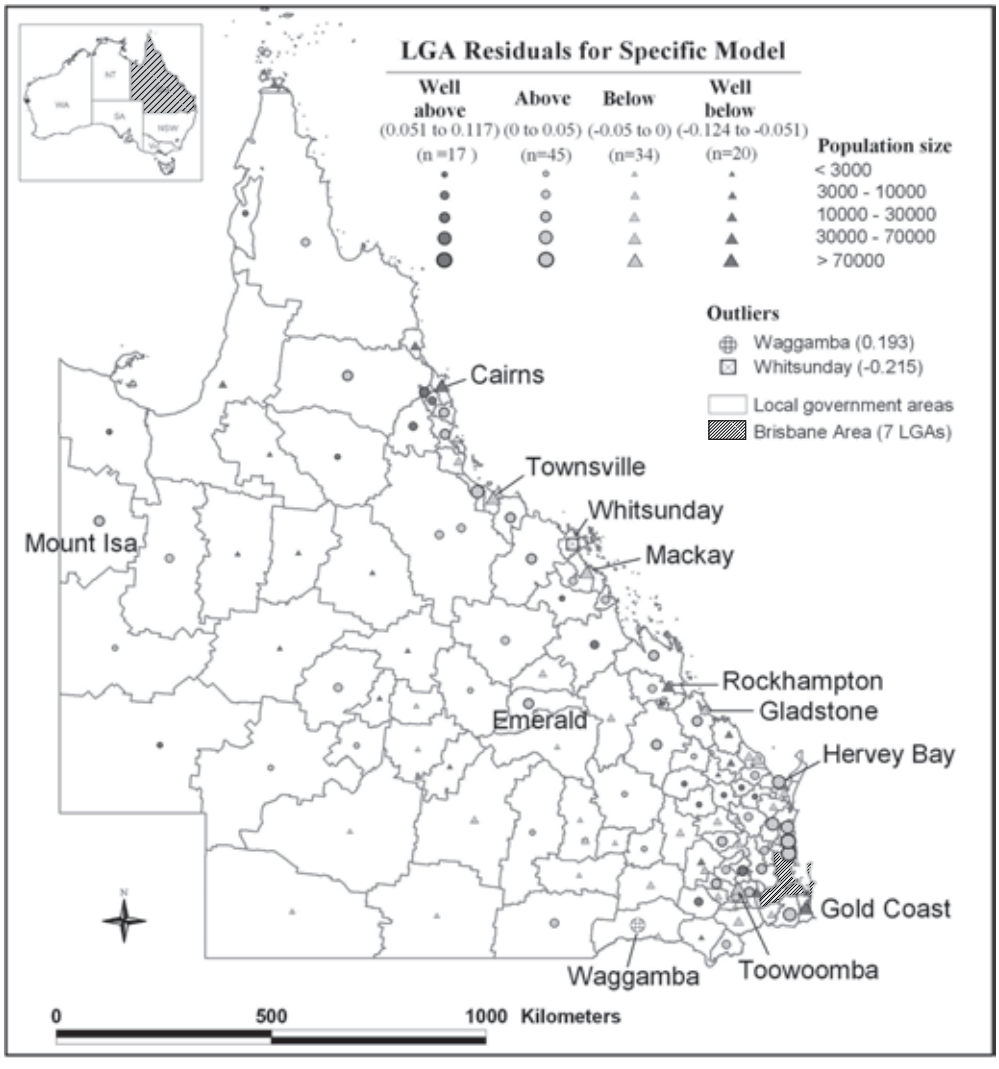
Source: Stimson et al. (2004).

Figure 18.2 Order of LGA residuals for the specific model

Figure 18.3 also identifies those LGAs for which the model estimates are well above or well below the regression line. There are 17 LGAs with residuals greater than +0.05, and these are located mainly around Cairns in the far north of Queensland and across the rural areas of the south-eastern part of the state inland from the coast. There are 20 LGAs with negative residuals greater than -0.05, and they include the coastal regional centres of Cairns and Rockhampton, the Gold Coast, some coastal rural places, and some isolated largely Indigenous places around the Gulf of Carpentaria and the Torres Strait islands.

Summary

The results of modelling conducted to identify the key factors that might explain the endogenous employment growth performance of LGAs across regional Queensland, using OLS regression and step-wise regression models (a general and a specific model), indicate that a reasonably small number of independent variables are important, with the model resulting in high R-squared values that explain over 88 per cent of the variance. The results tend to lend support to the importance of variables relating to industrial structure, population size and human capital as explanatory factors in explaining differential endogenous growth performance. Population growth emerges as a particularly strong positive factor, as is the level of income at the beginning of the decade period analysed. Population size at the beginning of the period is also important. But in addition, change in levels of unemployment as well as change in the concentration of occupational employment in Reich's routine production workers and in-person service workers, along with change in the concentration of employment in personal service, are also found to be significant determinants of regional endogenous employment growth/decline across Queensland's rural and regional LGAs. Finally, the modelling also identifies the significance of industrial specialization at the



Source: Stimson et al. (2004).

Figure 18.3 Spatial pattern of LGA residuals for the specific model

beginning of the period and of change in industrial specialization towards greater diversification, as processes that impact upon endogenous employment growth.

Future modelling approaches might include using alternative advanced techniques such as logit and probit models.

18.4 Towards a new paradigm for regional endogenous economic growth and development planning strategy

Because of the changing role of regional economies within nations and the impact of globalization, and given the context of contemporary concerns about how to achieve

sustainable development, a set of new considerations are now being taken into account in formulating and implementing economic development strategies for regions. These are outlined below:

1. Traditional models of regional economic growth and development and traditional modes of regional analysis remain important and useful as means of addressing regional economic change. But the emergence of the 'new growth theory' and the emphasis now being placed on endogenous processes in regional growth, along with the increasing concerns about the sustainability of development planning strategies, is resulting in the emergence of new integrated approaches to regional development policy and planning.
2. Many of the traditional approaches to strategy for regional economic development and the implementation of plans have been deficient – or at least appear to be inadequate – to deal with the dynamics of regions having both to compete in the global economy and to address issues of sustainability.
3. Following the influence of Porter (1985, 1986), increasingly it has become essential for regions to understand fully what factors constitute their regional competitiveness and how they might maintain and enhance that competitive position.
4. Authors such as Imbroscio (1995) advocate strategies of greater self-reliance, while Park (1995), McGee (1995) and Ohmae (1995) advocate the pursuit of regional economic development strategies based on strategic alliances and inter- and intra-regional network structures, including digital networks (Tapscott, 1996).
5. As indicated in the work of authors such as Henton (1995), Hall (1995), Waites (1995), Sternburg (1991) and Stough (1995), many regional analysts are advocating the need to base regional economic development on the growth of clusters of industries.
6. The concern over sustainability is increasingly evident in regional development and that is being reflected in regional plans which seek to integrate environmental, economic and social approaches to create urban and regional environments that enhance quality of life, meet environmental quality goals, and achieve economic growth and employment diversification.
7. In the contemporary era of globalization and an age of rapid change and uncertainty, procedures to identify and strategies to manage both exogenous and endogenous regional risks are crucial.

Thus, thinking is diverse on how to plan for and how to facilitate regional economic development in an environment of global competition, rapid change and a concern over sustainability.

All of this means that regions and regional economic development agencies need to give explicit attention not only to exogenous factors but also to endogenous factors in formulating regional development policy and in framing strategy and implementing plans to achieve regional economic growth and development, and to enhance regional performance. In arguing the need for an emerging paradigm of regional economic development planning, and to assist regions to undertake the processes involved in regional economic development strategy, it is important to identify the key elements for regional economic development strategy building and implementation, and to place those in a process that pulls together resources, infrastructure, social capital and technology to facilitate the

economic development of a region in a dynamic globally competitive environment. As suggested by Stimson et al. (2006), that includes giving explicit consideration to the following:

- the identification of regional core competencies, how to maintain them, and how to accumulate new core competencies;
- developing social capital;
- building and maintaining strategic leadership;
- the continuous rejuvenation or re-engineering of the processes of governance and the structure and functions of institutions;
- the more effective and efficient exploitation and management of resources;
- building market intelligence;
- providing strategic and smart infrastructure;
- identifying regional risks, and developing a risk management capability;
- incorporating the principles of sustainability into regional economic development strategies.

There is, however, no universal model or framework guaranteeing success for regional economic development. Stimson et al. (2006) propose a contemporary 'best-practice' approach to regional economic development. That approach suggests that the intent of regional economic development strategy might be to:

- establish a platform for change to guide the development of a region and to facilitate its competitiveness in a global environment in the pursuit of a sustainable future;
- mobilize key actors or facilitators and agents of change, through partnership approaches encompassing strategic alliances and partnerships between business, markets, government and community.

The framework proposes the following:

1. The identification, description, analysis and evaluation of core competencies, resource endowments, infrastructure competitiveness, market intelligence and regional risk through the combination of qualitative and quantitative methods encompassed in industry cluster analysis (ICA) and multi sector analysis (MSA) (see Roberts and Stimson, 1998).
2. The identification and evaluation of economic possibilities for the future, leading to the statement of strategic intent.
3. The evaluation of alternative development futures or scenarios through the participation of stakeholders within the region and, most importantly, external to it, to encompass the assessments of key decision-makers controlling capital, trade and other flows to the region. If the assessments of the feasibility of the alternative scenarios by the internal and external stakeholders are incongruent, then there is the potential that inappropriate or infeasible strategies will be pursued. Thus strategic directions might need to be redefined before formulating an economic development strategy which focuses on industry cluster development and the provision of strategic architecture.

4. Implementation plans and mechanisms need to be developed and put in place by appropriate agencies in the region.
5. The progress made towards achieving the desired development future needs to be monitored, requiring agreement on indicators and benchmarks set to measure and evaluate the performance of the region over time in order to assess the degree of success of the strategy and progress towards achievement of the strategic intent. Inevitably this involves building enhanced regional infrastructure systems along with strengthening existing and building new partnerships, networks and alliances.

Finally, in the contemporary era of the global economy, increasingly the pursuit of regional economic development also needs to take place within the context of principles for achieving a sustainable future.

Acknowledgement

The research on which this chapter is based is supported in part by a grant from the Australian Research Council, Discovery project # DP0558722.

Notes

1. For this chapter we define regional endogenous growth as the summation across all industry sectors of the differential/regional shift component derived from a shift-share analysis of employment change over a specified period of time, standardized by the size of a region's labour force. Whilst it is arguable exactly how to define regional endogenous growth, this definition has the advantage of being readily calculable using widely available census data and of being readily understood as a technique.
2. Particularly more recently to China and India.
3. Such as competition policy and tariff reductions.
4. While higher employment may appear to be desirable for a region, the value added by higher employment is also an important consideration. Growth in low value-added employment may produce other effects, such as housing affordability issues.
5. Derived from the Needleman technique (1978).
6. A p-value is the probability of obtaining a result at least as extreme, assuming the data point was the result of chance alone. Thus, the higher the p-value the greater the chance of it being obtained by chance alone.
7. Equivalent to a confidence interval of 95 per cent.
8. In some instances a p-value of marginally larger than 0.05 was included.

References

Anselin, T., A. Bera, R. Florax and M. Yoon (1996), 'Simple diagnostic tests for spatial dependence', *Regional Science and Urban Economics*, **26** (1), 77–104.

Arthur, W.B. (1994), *Increasing Returns and Path Dependency in the Economy*, Ann Arbor, MI: University of Michigan Press.

Barro, R.J. (1990), 'Endogenous technological change', *Journal of Political Economy*, **98**, S71–S102.

Bradley, R. and J. Gans (1998), 'Growth in Australian cities', *Economic Record*, **74** (226), 266–78.

Bureau of Transport and Regional Economics (2004a), 'Focus on Regions No. 1: Industry Structure', Information Paper 49, BTRE, Department of Transport and Regional Services, Canberra.

Bureau of Transport and Regional Economics (2004b), 'Focus on Regions No. 2: Education, Skills and Qualifications', Information Paper 51, BTRE, Department of Transport and Regional Services, Canberra.

Cass, D. (1965), 'Optimal growth in an aggregate model of capital accumulation', *Review of Economic Studies*, **32**, 233–40.

Castells, M. and P. Hall (1994), *Technopoles of the World: The Making of 21st Century Industrial Complexes*, London: Routledge.

Draca, M., J. Foster and C. Green (2003), 'Human capital investment and economic growth in the Australian economy', *Productivity and Regional Economic Performance in Australia*, Brisbane: Queensland Office of Economic and Statistical Research (OESR).

Duranton, G. and D. Puga (2000), 'Diversity and specialisation in cities: why, where and when does it matter?', *Urban Studies*, **37** (3), pp. 533–55.

Erickson, R.A. (1994), 'Technology, industrial restructuring and regional development', *Growth and Change*, **25**, 353–97.
Erickson, R.A. and T. Leinbach (1979), 'Characteristics of branch plants attracted to non metropolitan areas', in R. Lonsdale and H.L. Seyter (eds), *Non Metropolitan Industrialisation*, Washington, DC: Winston.
Fukuyama, F. (1995), *Trust: The Social Virtues and Creation of Prosperity*, New York: Free Press.
Garlick, S., M. Taylor and P. Plummer (2006), *An Enterprising Approach to Regional Growth: The Role of VET in Regional Development*, NCVER (National Council for Vocational Education Research), Australian National Training Authority Adelaide.
Garnaut, J., P. Connell, R. Lindsay and V. Rodriguez (2001), 'County Australia: influences on employment and population growth', *ABARE Research Report*, 2001.1, Canberra.
Garnett, A.M. and E.T. Lewis (2000), 'Population and labour mobility in rural Australia', *Australasian Journal of Regional Studies*, **6** (2), 157–72.
Goetz, S.J. and A. Rapasingla (2001), 'The returns to higher education: estimates for the contiguous states', paper to Regional Science Association International, North American Annual Meeting, Charleston, SC, November.
Gordon, I.R. and P. McCann (2000), 'Industrial clusters, complexes, agglomeration and/or social networks', *Urban Studies*, **37** (3), 513–32.
Grossman, G.M. and E. Helpman (1991), *Innovation and Growth in the Global Economy*, Cambridge, MA: MIT Press.
Hall, P. (1995), 'The roots of urban innovation: culture, technology and urban order', in *Cities and the New Global Economy*, An international conference presented by the OECD and the Australian Government, Melbourne, November 1994, Canberra: Australian Government Publishing Service, pp. 275–93.
Hanushek, E.A. and D.D. Kimko (2000), 'Schooling, labour-force quality, and the growth of nations', *American Economic Review*, **90** (5), pp. 1884–1208.
Harrison, H. (1997), 'Trends in the delivery of rural health, education and banking services', *National Focus,* National Farmers Federation Research Paper, vol. II, Canberra, February.
Haynes, K.E. and M. Dinc (1997), 'Productivity change in manufacturing regions: a multi-factor shift share approach', *Growth and Change*, **28**, 201–21.
Henderson, J.V., A. Kuncoro and M. Turner (1995), 'Industrial development in cities', *Journal of Political Economy*, **103**, 1067–90.
Henton, D. (1995), 'Reinventing Silicon Valley: creating a total quality community', in *Cities and the New Global Economy*, An international conference presented by the OECD and the Australian Government, Melbourne, November 1994, Canberra: Australian Government Publishing Service, pp. 306–26.
Imbroscio, D. (1995), 'An alternative approach to urban economic development: exploring the dimensions and prospects', *Urban Affairs*, **30**, 840–67.
Johansson, B., Ch. Karlsson and R.R. Stough (2001), *Theories and Endogenous Growth: Lessons for Regional Policy*, Heidelberg: Springer-Verlag.
Kaufman, R. (1993), 'An empirical exploration of the relation among diversity, stability and performance in economic systems', *Structural Change and Economic Dynamics*, **4** (2), 299–313.
Koopmans, T.C. (1965), 'On the concept of optimal economic growth', *The Econometric Approach to Development Planning*, Pontificae Academiae Scientiarum Scripta Varia. **28**, pp. 225–300.
Krugman, P. (1991), *Geography and Trade*, Cambridge, MA: MIT Press.
Lande, P. (1994), 'Regional industrial structure and economic growth and stability', *Journal of Regional Science*, **34** (3), 343–60.
Lawson, J. and J. Dwyer (2002), *Labour Market Adjustment in Regional Australia*, Sydney: Reserve Bank of Australia.
Lucas, R.E. (1985), *Models of Business Cycles*, Oxford: Basil Blackwell.
Lucas, R.E. (1988), 'On the mechanics of economic development', *Journal of Monetary Economics*, **22**, 3–42.
McGee, T.G. (1995), 'System of cities and networked landscapes: new cultural formations and urban built environments in the Asia-Pacific region', Pacific Rim Council on Urban Development, Conference Proceedings Brisbane.
Needleman, L. (1978), 'On the approximation of the Gini coefficient of concentration', *Manchester School*, **46**, 106–22.
Ngai, R.L. (2004), 'Barriers and the transition to modern growth', *Journal of Monetary Economics*, **51** (7), 1353–83.
Norris, K. and M. Wooden (1996), *The Changing Australian Labour Market*, Canberra: Australian Government Publishing Service.
Norton, R.D., and J. Rees (1979), 'The product cycle and the decentralization of North American manufacturing', *Regional Studies*, **13**, 141–51.
Ohmae, K. (1995), *The End of the Nation State: The Rise of Regional Economics*, New York: Free Press.
Parente, S.L., (2001), 'The failure of endogenous growth', *Knowledge Technology and Policy*, **13** (4), 49–58.

Parente, S.L. and E.C. Prescott (1994), 'Barriers to technology adoption and development', *The Journal of Political Economy*, **102** (2), 298–321.
Park, S.O. (1995), 'Networks and competitive advantages of new industrial districts', paper presented to the Pacific Regional Science Conference Organisation, 14th biennial meeting, July, Taipai.
Porter, M.E. (1985), *Competitive Advantage: Creating and Sustaining Superior Performance*, New York: Free Press.
Porter, M.E. (1986), *Competition in Global Industries*, Boston, MA: Harvard Business School Press.
Porter, M.E. (1990), *The Competitive Advantage of Nations*, New York: Macmillan.
Productivity Commission (1998), 'Aspects of structural change in Australia', Research Report, Ausinfo, Canberra.
Rebelo, S. (1991), 'Long run policy analysis and long run growth', *Journal of Political Economy*, **98**, S71–S102.
Reich, R. (1991), *The Work of Nations: Preparing Ourselves for 21st Century Capitalism*, New York: Vintage Books.
Rees, J. (1979), 'State technology programs and industry experience in the USA', *Review of Urban and Regional Development Studies*, **3**, 39–59.
Rees, J. (2001), 'Technology and regional development: theory revisited', in B. Johansson, Ch. Karlsson and R.R. Stough (eds), *Theories of Endogenous Regional Growth*, Heidelberg: Springer-Verlag, pp. 94–110.
Roberts, B.H. and R.J. Stimson (1998), 'Multi-sectoral qualitative analysis: a tool for assessing the competitiveness of regions and developing strategies for economic development', *Annals of Regional Science*, **32** (4), 459–67.
Romer, P. (1986), 'Increasing returns and long run growth', *Journal of Political Economy*, **94**, 1002–37.
Romer, P.M. (1990), 'Endogenous technological change', *Journal of Political Economy*, **98**, S71–S102.
Saxenian, A. (1994), *Regional Advantage: Culture, and Competition in Silicon Valley and Route 128*, Cambridge, MA: Harvard University Press.
Scott, A.J. (1988), *New Industrial Spaces: Flexible Production Organization and Regional Development in North America and Western Europe*, London: Pion.
Simmie, J. (ed.) (1997), *Innovation, Networks and Learning Regions?*, London: Jessica Kingsley.
Solow, R.M. (1956), 'A contribution to the theory of economic growth', *Quarterly Journal of Economics*, **70**, 65–94.
Solow, R.M. (2000), *Growth Theory: An Exposition*, New York: Oxford University Press.
Sternburg, E. (1991), 'The sectoral clusters in economic development policy: lessons from Rochester and Buffalo', *Economic Development Quarterly*, **4**, 342–56.
Stimson, R., A. Robson and T.-K. Shyy (2004), 'Shift-share analysis and modeling endogenous growth across Queensland's regions', *North American Regional Science Council, 52nd Annual Meeting*, Seattle, November.
Stimson, R., R. Stough and B. Roberts (2006), *Regional Economic Development: Analysis and Planning Strategy*, Berlin: Springer.
Stimson, R., R. Stough and M. Salazar (2005), 'Leadership and institutional factors in endogenous regional economic development', *Investigaciones Regionales*, **7**, 23–52.
Stimson, R., S. Baum, J. Mangan, Y. Van Gellecum, T.-K. Shyy and T. Yigitcanlar (2004), 'Analysing spatial patterns in the characteristics of work, employment and skills in Australia's capital cities and across its regional cities and towns: modelling community opportunity and vulnerability, main report', prepared for the Australian National Training Authority (ANTA) 2003 National Project, Centre for Research into Sustainable Urban and Regional Futures, University of Queensland, Brisbane.
Stough, R.R. (1995), 'Industry sector analysis of the Northern Virginia Economy', *Proceedings of the Second Annual Conference on the Future of the Northern Virginia Economy*, Center for Regional Analysis, The Institute of Public Policy and the Northern Virginia Business Roundtable, George Mason University, Fairfax, VA, 3–37.
Tapscott, D. (1996), *The Digital Economy: Promise and Peril in the Age of Networked Intelligence*, New York: McGraw Hill.
Taylor, P.J., G. Catalano and N. Gane (2002), 'A geography of global change: services and cities 2000–01', *GaWC Research Bulletin*, Globalisation and World Cities Study Group and Network, **77**, 1–9.
Thomas, M.D. (1975), 'Growth pole theory, technology change and regional economic growth', *Papers of the Regional Science Association*, **34**, 3–25.
Waites, M. (1995), 'Economic development: building and economic future', *State Government News*, **38**.

19 Regional growth and convergence: heterogeneous reaction versus interaction in spatial econometric approaches

Cem Ertur and Julie Le Gallo

> Notwithstanding the general rule that 'everything affects everything else', it is often useful to assess whether the dominant effects are caused by *reaction* to external forces or by *interaction* between (neighbouring) individuals. (Cliff and Ord, 1981, p. 141)

19.1 Introduction

Over the last few years, numerous studies have been carried out to analyze economic convergence among countries or regions, recognizing at the same time the need to include spatial effects (Abreu et al., 2005; Ertur et al., 2006; Fingleton and López-Bazo, 2006). For example, a large number of contributions analyzing the β-convergence hypothesis impose strong homogeneity assumptions on the cross-economy growth process, since each economy is assumed to have an identical aggregate production function. However, modern growth theory suggests that different economies should be described by distinct production functions. In other words, β-convergence models should account for parameter heterogeneity (Brock and Durlauf, 2001; Durlauf, 2001; Durlauf et al., 2005; Temple, 1999). Evidence of parameter heterogeneity has been found in non-spatial models using different statistical methodologies, such as in Canova (2004), Desdoigts (1999), Durlauf and Johnson (1995) and Durlauf et al. (2001). Each of these studies suggests that the assumption of a single linear statistical growth model applying to all countries or regions is incorrect.

Moreover, Ertur et al. (2007) argue that in a spatial context, similarities in legal and social institutions, as well as culture and language, might create spatially local uniformity in economic structures, leading to situations where rates of convergence are similar for observations located nearby in space. Parameter heterogeneity is then spatial in nature and estimating a 'global' relationship between growth rate and initial per capita income, which applies in the same way over the whole study area, does not allow for capturing the important convergence rate differences that might occur in space.

The instability in space of economic relationships illustrated by this example is called spatial heterogeneity. This phenomenon can be observed at several spatial scales: behaviors and economic phenomena are not similar in the center and in the periphery of a city, in an urban region and in a rural region, in the 'West' of the enlarged European Union and in the 'East', and so on. In an econometric regression, these differences may appear in two ways: with space-varying coefficients and/or space-varying variances. The first case is labeled structural instability of regression parameters, which vary systematically in space. The second case pertains to heteroskedasticity, which is a frequent problem in cross-sections.

Spatial heterogeneity is one of the two spatial effects analyzed by the field of spatial econometrics (Anselin, 1988). This effect operates through the specification of the

reaction of the variable of interest to explanatory variables, that is through the specification of its conditional mean using a spatially varying parameter scheme, or the specification of its conditional variance imposing a known spatial structure on the variance-covariance matrix of the error term. The other is spatial autocorrelation, or the coincidence of value similarity and locational similarity. It is aimed at capturing interaction between neighboring units of observation. This effect is also highly relevant in growth and convergence analysis. Indeed, as pointed out by Easterly and Levine (2001), there is a tendency for all factors of production to gather together, leading to a geographic concentration of economic activities. As a consequence, any empirical study on growth and convergence should explicitly acknowledge this phenomenon of spatial interdependence between regions or countries. Moreover, as pointed out by Abreu et al. (2005), this distinction between spatial heterogeneity and spatial dependence can be related to two different ways of modeling spatial data in growth regressions: models of absolute location and models of relative location. Absolute location refers to the impact of being located at a particular point in space (continent, climate zone) and is usually captured through dummy variables. Relative location refers to the effect of being located closer or further away from other specific countries or regions.

While spatial autocorrelation has been the focus of several literature reviews (Anselin and Bera, 1998; Anselin, 2006 for instance), spatial heterogeneity is much less presented per se. In the convergence context, spatial effects have also already been the focus of several literature reviews: Abreu et al. (2005), Rey and Janikas (2005) and Fingleton and López-Bazo (2006). However, these studies focus more on the appropriate treatment and interpretation of spatial autocorrelation in convergence models and/or distribution dynamic approaches. Abreu et al. (2005) present some models of absolute location but limit their discussion to models for discrete spatial heterogeneity, while several recent studies extend these to models with continuous space-varying coefficients (Bivand and Brunstad, 2005; Eckey et al., 2007; Ertur et al., 2007).

In this context, this chapter has two aims. Firstly, we present the main econometric specifications capturing spatial heterogeneity, or models of absolute locations in the terminology of Abreu et al. (2005). Here, we focus on structural instability, as well as on specific forms of heteroskedasticity and we provide examples of applications pertaining to growth econometrics. Secondly, we examine how these specifications can be extended to allow further for spatial autocorrelation in models of heterogeneous reaction. Concerning this second point, it should be noted that spatial autocorrelation and spatial heterogeneity entertain complex links. Firstly, as pointed out by Anselin and Bera (1998) and Abreu et al. (2005), there may be observational equivalence between these two effects in a cross-section. Indeed, a cluster of high-growth regions may be the result of spillovers from one region to another, or it could be due to similarities in the variables affecting the regions' growth. Secondly, heteroskedasticity and structural instability tests are not reliable in the presence of spatial autocorrelation. For instance, Anselin and Griffith (1988) show that spatial autocorrelation affects the size and power of the White and Breusch–Pagan tests of heteroskedasticity. Anselin (1990a) also provides evidence that the Chow test for structural instability is not reliable in the presence of spatial autocorrelation. Conversely, spatial autocorrelation tests are affected by heteroskedasticity (Anselin, 1990b). Thirdly, spatial autocorrelation is sometimes the result of an unmodeled parameter instability (Brunsdon et al., 1999a). In other words, if space-varying relationships

are modeled within a global regression, the error terms may be spatially autocorrelated. We detail some of these issues in the growth and convergence context in this chapter.

Note that heterogeneity can also be modeled using spatial panel data models. However, this alternative approach will not be considered here as a complete survey is provided by Anselin et al. (2008). Therefore, we focus here exclusively on the cross-sectional approach. Also, due to space constraints, we do not review distribution dynamics approaches to economic convergence and focus exclusively on spatial econometric modeling issues.[1] Bearing these different elements in mind, this chapter is organized as follows. The next section presents the specifications allowing for discrete heterogeneity, that is, when different parameters are estimated following spatial regimes. The following sections are devoted to continuous heterogeneity models: geographically weighted regressions (section 19.3) and their generalizations (section 19.4). Section 19.5 concludes and provides some research directions.

19.2 Discrete spatial heterogeneity

Spatial instability of the parameters necessitates specifications in which the characteristics of each spatial observation are taken into account. Therefore, we could specify a different relationship for each zone i of the sample:

$$y_i = x_i'\beta_i + \varepsilon_i \qquad i = 1,\ldots N \tag{19.1}$$

where y_i represents the observation of the dependent variable for zone i; x_i' is the $(1,K)$ vector including the observations for the K explanatory variables for zone i. It is associated to β_i, a $(K,1)$ vector of parameters to be estimated. Finally, in general, the variance differs with i: $\varepsilon_i \sim iid(0,\sigma_i^2)$. Of course, given N observations, it is not possible to estimate consistently NK parameters and N variances: this is the incidental parameter problem. Therefore, a spatial structure for the data must be specified. The spatial variability of the mean of the coefficients of a regression can be discrete, if systematic differences between regimes are observed, or it can be continuous over the whole area. Note that, similarly to panel data models, a random variation could also be specified, under the form of a random coefficients model. As this possibility is not explicitly spatial, it will not be further considered in this chapter.[2]

Consider first the models for discrete spatial heterogeneity, which have been applied extensively to study the club convergence hypothesis in a spatial context. Assume that the area under study is divided into several regimes. If only one variable is under study, a spatial ANOVA (Analysis of Variance) can be undertaken in order to investigate whether the mean of this variable is different across the regimes. Spatial versions of ANOVA have also been suggested by Griffith (1992).

More generally, in a regression model, consider the case of two regimes, indicated by 1 and 2.[3] It can be written as follows:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \tag{19.2}$$

where y_1 and y_2 are the $(N_1,1)$ and $(N_2,1)$ vectors of observations for the dependent variable; X_1 and X_2 are the (N_1,K) and (N_2,K) matrices of observations of the explanatory variables; β_1 and β_2 are the unknown vectors of parameters to be estimated. Let

$\varepsilon' = [\varepsilon_1' \; \varepsilon_2']$ be the vector of error terms. If $E[\varepsilon\varepsilon'] = \sigma^2 I_N$, the test of spatial homogeneity $\beta_1 = \beta_2$ can be performed with the traditional Chow test. However, more sophisticated error structures can be specified, such as groupwise heteroskedasticity (19.3) and/or spatial error autocorrelation (19.4):

$$\Psi = E[\varepsilon\varepsilon'] = \begin{bmatrix} \sigma_1^2 I_{N_1} & 0 \\ 0 & \sigma_2^2 I_{N_2} \end{bmatrix} \tag{19.3}$$

$$\varepsilon = \rho W\varepsilon + u \quad u \sim iid(0,\sigma_u^2) \tag{19.4}$$

where W is a (N,N) spatial weights matrix. Both possibilities can be combined or a different spatial process may be specified for each regime. In each case, maximum likelihood should be carried out and the Chow test must be spatially adjusted (Anselin, 1990a).

This framework has been applied to consider specific forms of parameter heterogeneity in absolute β-convergence regressions, in which case the explanatory variable is the growth rate of per capita income and the explanatory variable is the initial per capita income. Indeed, while absolute β-convergence is frequently rejected for large sample of countries and regions, it is usually accepted for more restricted samples of economies belonging to the same geographical area. This observation can be linked to the presence of convergence clubs: there is not only one steady state to which all economies converge. From an econometric point of view, one equation must be estimated for each club and decisions must be made as to how the cross-sectional sample should be partitioned.

Some papers just use a priori spatial regimes, such as Northern and Southern European regions (Neven and Gouyette, 1995), or regions belonging to cohesion countries and the others (Ramajo et al., 2008). Exploratory spatial data analysis may also prove useful in this task. Indeed, these techniques, by exploiting the specific spatial nature of the data, are useful in characterizing the form of spatial heterogeneity by detecting the local concentrations of similar values, by using Getis–Ord statistics (Ord and Getis, 1995) or LISA (Local Indicators of Spatial Association) statistics (Anselin, 1995). For example, Le Gallo and Ertur (2003) show that the spatial distribution of per capita gross domestic product (GDP) in Europe before the recent enlargement in 2003 is characterized by a strong North–South polarization. More recently, Ertur and Koch (2006a) show that this polarization scheme is replaced by a new West–East polarization scheme if the last enlargement of the European Union to include Central and Eastern European countries is taken into account. These polarization schemes represent evidence in favor of the existence of at least two spatial regimes in the European regions. This information is then used to estimate β-convergence models with spatial regimes as in equation (19.2) (Fischer and Stirböck, 2006; Le Gallo and Dall'erba, 2006), possibly associated with groupwise heteroskedasticity, and spatially autocorrelated error terms as in equations (19.3) and (19.4) (Ertur et al., 2006), or a spatial lag of the form Wy (Dall'erba and Le Gallo, 2008). However, as pointed out by Rey and Janikas (2005), the existing specification search procedures should be extended to be able to distinguish between spatial dependence and spatial heterogeneity while formal specification search strategies for spatial heterogeneity have yet to be suggested.

While non-spatial papers use endogenous detection methods, such as regression trees (Durlauf and Johnson, 1995), it should be emphasized that a technique allowing for an endogenous estimation of regimes together with taking into account of spatial

autocorrelation stills needs to be developed (Anselin and Cho, 2002). A first step in this direction is the paper by Basile and Gress (2005) who suggest a semi-parametric spatial autocovariance specification that simultaneously takes into account the problems of non-linearities and spatial dependence. In that purpose, they extend Liu and Stengos' (1999) non-parametric specification by allowing a spatial lag term or a spatial error process.

If no information is available on spatial regimes, or if one thinks that the mean of a variable or that the regression coefficients do not change brutally between regimes, it is preferable to use specifications allowing for continuous spatial variations across the whole study area. The urban literature has frequently used trend surface analysis models and/or the expansion method. In the first case, the coordinates of each location (such as latitude and longitude) are added in the regression model so that the main characteristics of the regression surface, such as simple 'North–South' or 'East–West' drifts or more complex drifts for higher-order functions, can be described (Agterberg, 1984). In the second case, the regression coefficients are deterministic (Casetti, 1972) or stochastic (Anselin, 1988) functions of expansion variables, such as the coordinates of each location. However, the expansion method suffers from two main drawbacks (Fotheringham et al., 2000, 2004). Firstly, these techniques only allow for capturing trends in relations in space, the complexity of these trends being determined by the complexity of the specified expansion equations. The estimates of the parameters may therefore obscure important local variations to the broad trends represented by the expansion equations. Secondly, the form of the expansion equations must be specified a priori. To overcome these problems, geographically weighted regression (GWR) has been developed and applied in several papers focusing on economic convergence.

19.3 Geographically weighted regression (GWR)

The geographically weighted regression (GWR), or equivalently the locally linear regression method LWR (Locally Weighted Regression), has been developed by McMillen (1996) and Brunsdon et al. (1996). Most details concerning this method are developed in two books (Fotheringham et al., 2000, 2004). GWR is a locally linear, non-parametric estimation method aimed at capturing, for each observation, the spatial variations of the regression coefficients. For that purpose, a different set of parameters is estimated for each observation by using the values of the characteristics taken by the neighboring observations.

Formally, consider again as a point of departure the general formulation (19.1) where a vector of K unknown parameters must be estimated for each observation i:

$$y_i = x_i'\beta_i + \varepsilon_i = \sum_{k=1}^{K} \beta_{ik} x_{ik} + \varepsilon_i \tag{19.5}$$

where $\varepsilon_i \sim iid(0,\sigma^2)$, $i = 1,...,N$. In order to estimate the parameters β_{ik} of model (19.5), we assume that observations close to location i exert more influence on the estimation of β_{ik} than those located farther away. The idea is then to use a distance-decay weighting scheme that spatially varies with i. Formally, let β_i be the weighted least squares (WLS) estimator of the vector β_i of the K unknown parameters. It is written in matrix form as:

$$\hat{\beta}_i = (X' V_i X)^{-1} X' V_i y \tag{19.6}$$

with the same notations as before and $V_i = diag[v_{i1}, v_{i2}, ..., v_{iN}]$ is a (N,N) diagonal matrix, specific to each location i. The diagonal elements of V_i represent the geographic weighting given to the observations surrounding i, generally specified using a continuous and monotone decreasing function of the distance between location i and all other observations, in other words a kernel function.

This methodology differs from the traditional non-parametric kernel estimation where the weights refer to the attribute space of the explanatory variables (Cleveland et al., 1988). In contrast, GWR uses weights referring to the location in geographical space and therefore allows for estimating local rather than global parameters. Different weighting schemes or kernel functions have been suggested in the literature (McMillen, 1996; MacMillen and McDonald, 1997; Fotheringham et al., 2000). One of the most commonly used weighting function is the Gaussian kernel, for a given location i, we have:

$$v_{ij} = \exp(-d_{ij}^2/h^2) \quad j = 1,...,N \tag{19.7}$$

where d_{ij} is the Euclidian distance between locations i and j and h is referred to as the bandwith parameter that can be determined by a cross-validation procedure. Another possibility is to use a truncated kernel by setting the weights to zero outside a radius d and to decrease monotonically to zero inside the radius as d_{ij} increases. For example consider a bisquare weighting function written as:

$$v_{ij} = \begin{cases} (1 - d_{ij}^2/d^2)^2 & \text{if } d_{ij} \leq d, \\ 0 & \text{if } d_{ij} > d \end{cases} \tag{19.8}$$

or even a tri-cube weighting function as suggested by McMillen (1996) and McMillen and McDonald (1997):

$$v_{ij} = \left[1 - \left(\frac{d_{ij}}{d_i}\right)^3\right]^3 I(d_{ij} < d_i) \tag{19.9}$$

where d_i is the distance of the mth nearest observation to i and $I(\cdot)$ is an indicator function that equals one when the condition is true. The window size, m, is the number of nearest neighbors and determines the observations which receive non-zero monotonically decreasing weights, whereas the observation farther away are given zero weights. Again it can be determined by cross-validation.

Note also that a mixed version of GWR has been suggested by Brunsdon et al. (1999a) and Mei et al. (2004, 2006), in which some coefficients are allowed to vary in space while others remain constant. From an empirical point of view, GWR is useful to identify the nature and patterns of spatial non-stationarity over the studied area. Indeed, the result of a GWR is a set of localized estimations of the parameters, together with localized versions of t-statistics and measures of quality of fit. These local measures are associated to specific locations, so that they can be mapped to illustrate the spatial variations of the relationship under study (Mennis, 2006).

We review here some of the most recent contributions related to regional growth and development. Bivand and Brunstad (2005), in their paper focusing on the detection of spatial misspecification in growth models using the R software, estimate a conditional

convergence model including a spatially lagged endogenous variable and spatial regimes for Western Europe over the period 1989–99. They find support for the role of agricultural subsidies in accounting for variations in regional growth. Higher levels of agricultural support are associated with lower levels of growth, even after some measure of human capital has been introduced. They also consider a GWR specification essentially to ascertain their results by exploring whether any traces of remaining spatial non-stationarity can be found. However, they do not fully interpret their GWR results due to some methodological problems which will be discussed below.

Another attempt to use GWR regressions in the regional growth context has been made by Eckey et al. (2007) in a paper focusing on regional convergence in Germany over the period 1995–2002, using disaggregated data on a sample of 180 labor market regions. They estimate a model based on Mankiw et al. (1992) allowing all coefficients, especially the rate of convergence, to vary across regions. Each region seems to converge using both absolute and conditional convergence models as the local convergence parameters are all negative. The value of the convergence speed increases from south to north. The half-life period ranges from less than 20 years for some regions in northern Germany to more than 50 years for regions in southern Bavaria.

Finally, let us mention the contribution of Yu (2006) to the regional development literature in his study of the development mechanisms in the Greater Beijing Area using GWR. The analysis reveals two results: first, regional development mechanisms in the Greater Beijing Area, such as foreign direct investment, per capita fixed asset investment and percentage of fixed assets invested in state-owned enterprises, show significant spatial non-stationarity; and second, development mechanisms have strong local characteristics.

From a methodological point of view, several problems plaguing GWR estimation and inference must be mentioned here. Firstly, concerning statistical inference, in order to know whether the local estimations of parameters are significantly different between them and compared to the OLS estimator, parametric tests have been suggested by Brunsdon et al. (1999a) and Leung et al. (2000a). Secondly, LeSage (2004) argues that the presence of aberrant observations due to spatial enclave effects, shifts in regime or outliers can exert undue influence on the GWR estimates. Therefore, he suggests a Bayesian estimation approach that detects these observations and down-weights them to lessen their influence on the estimates. Thirdly, Wheeler and Tiefelsdorf (2005) point out that the local regression estimates are potentially collinear even if the underlying exogenous variables in the data-generating process are uncorrelated. This collinearity can degrade coefficient precision in GWR and lead to counter-intuitive signs for some regression coefficients. Using Monte Carlo simulations, Wheeler and Calder (2007) show that Bayesian models with spatially varying coefficients (Gelfand et al., 2003) provide more accurate regression coefficients. Finally, facing the various inference problems encountered by GWR, Páez et al. (2002a) place GWR in a different statistical framework, interpreting GWR as a spatial model of error variance heterogeneity, that is, heteroskedasticity. The variance of the error term is defined as an exponential function of the squared distance between two observations and has then a precise geographical interpretation. While this approach is a special case of the well-known multiplicative heteroskedasticity model developed by Harvey (1976), it nevertheless represents a real breakthrough in the GWR literature and allows the derivation of formal heterogeneity tests.

There still remains an important methodological problem pointed out by Páez et al. (2002b) and Pace and LeSage (2004): spatial dependence may not be eliminated even at the optimal bandwidth as is often assumed in the related literature, where it is considered that spatial dependence is mainly due to inadequately modeled spatial heterogeneity. Actually, this methodological problem is related to the complex links between spatial heterogeneity and spatial dependence often underlined, and more generally to the reaction versus interaction debate first pointed out by Cliff and Ord (1981, p. 141) in the spatial econometrics literature. Even when heterogeneous reactions are taken into account as in the GWR framework, it could be the case that there are also interactions between units of observation that should be modeled with a spatially dependent covariance structure. Therefore, Brunsdon et al. (1998) have proposed to include the spatially lagged endogenous variable in the GWR model and Leung et al. (2000b) have suggested a test of spatial autocorrelation of the GWR residuals. Moreover, Páez et al. (2002b) formulate a general model of spatial effects that includes as special cases GWR with a spatially lagged endogenous variable (GWR-SL) and GWR with spatially autocorrelated residuals (GWR-SEA). Finally, Pace and LeSage (2004) introduce spatial autoregressive local estimation (SALE) based on a computationally competitive recursive maximum likelihood estimation method.

19.4 Generalized GWR

We first consider here a straightforward generalization of the model proposed by Páez et al. (2002b) where the spatial lags of the explanatory variables are also added in the model:

$$\begin{cases} y = \rho W_1 Y + \tilde{X}\beta + \varepsilon \\ \varepsilon = \lambda W_2 \varepsilon + u \end{cases} \tag{19.10}$$

where $u \sim N(0,\Omega)$; y is the $(N,1)$ vector of the dependent variable; $\tilde{X} = [\iota \ \underline{X} \ W_1 \ \underline{X}]$ with τ a $(N \times 1)$ unit vector; $\underline{X}$ a $(N,(K-1))$matrix of the explanatory variables excluding the constant and $W_1\underline{X}$ its spatial lag; β is the $((2(K-1)+1),1)$ vector of the associated parameters to be estimated; ρ and λ are the spatial autoregressive parameters; W_1 and W_2 are row-standardized spatial weights matrices; Ω is the diagonal covariance matrix of the error term u with elements denoted by ω_{ii}. More precisely, they adopt a specific form for this covariance matrix as follows: $\Omega = \sigma^2 G$ and define its elements as $\omega_{ii} = \sigma^2 g_i(\gamma, z_i)$ and $\omega_{ij} = 0$ for $i \neq j$. Hence the variance of the error term u is a function of a $(p,1)$ vector of known variables z_i, an unobservable parameter vector γ and an unknown constant σ^2. The geographically weighted specification is then obtained by defining a variance model of the exponential form as in Páez et al. (2002a):

$$g_{oi}(\gamma_o, a_{oi}) = \exp(\gamma_o d_{oi}^2) \tag{19.11}$$

which is a special case of the previous formulation with $p = 1$ and where the observable variable d_{oi} is the distance between location o and observation i for $i = 1,...,N$. This particular geographical specification of the error variance is called locational heterogeneity by Páez et al. (2002a, 2002b). The parameter γ_o is then the so-called kernel bandwidth in

the tradition GWR literature. The generalized GWR model can therefore be defined in terms of local parameters as follows:

$$\begin{cases} y = \rho_o W_1 y + \tilde{X}\beta_o + \varepsilon_o \\ \varepsilon_o = \lambda_o W_2 \varepsilon + u_o \end{cases} \Rightarrow \begin{cases} A_o y = \tilde{X}\beta_o + \varepsilon_o \\ B_o \varepsilon_o = u_o \end{cases} \tag{19.12}$$

where $A_o = I - \rho_o W_1$; $B_o = I - \lambda_0 W_2$ and $u_o \sim N(0,\sigma_o^2 G_o)$. Note that A_o and B_o depend on local parameters ρ_o and λ_o respectively and G_o depends on the local parameter γ_o.

If no spatial lags of the explanatory variables are allowed, that is, $\tilde{X} = [\iota \ \underline{X}]$, it is easily seen that when $\rho_o = \lambda_o = 0$ then $A_o = B_o = I$, and this model reduces to the standard GWR model; when $\lambda_o = 0$ then $B_o = I$, and we obtain a GWR model which includes the spatially lagged endogenous variable (GWR-SL); when $\rho_o = 0$ then $A_o = I$, and we obtain a GWR model with spatially autocorrelated errors (GWR-SEA). Páez et al. (2002b) propose to estimate those two generalized GWR models by iterated maximum likelihood. They also derive formal Lagrange multiplier tests against several forms of misspecification including a test for omitted endogenous spatial lag, a test for spatial error autocorrelation in GWR models and tests for locational heterogeneity in global models in the presence of a spatially lagged endogenous variable or in the presence of spatial error autocorrelation. More flexibility is allowed in the specification of the model by using $\tilde{X} = [\iota \ \underline{X} \ W_1 \ \underline{X}]$ which also includes spatial lags of the explanatory variables; the estimation method as well as all of the tests proposed by Páez et al. (2002b) may then be straightforwardly generalized to such a model at practically no cost.

An alternative approach to the generalization of the GWR model is proposed by Pace and LeSage (2004): spatial autoregressive local estimation (SALE) allows for simultaneously considering spatial parameter heterogeneity and spatial autocorrelation in an efficient way using recursive spatial maximum likelihood. Their approach is based on the estimation of a sequence of N spatial autoregressions, one for each observation, using a range of sub-sample sizes. Consider the spatial Durbin Model (SDM) where the spatial lags of the explanatory variables are also added in the model:

$$y = \tilde{X}\beta + \rho W y + \varepsilon \tag{19.13}$$

where the same notations as before and assuming that $\varepsilon \sim N(0,\sigma^2 I_N)$. The concentrated log-likelihood function for the global SDM model is then written as follows, for fixed ρ, omitting the constant term (Pace and Barry, 1997):

$$L(\rho) = \ln\left|I - \rho W\right| - \frac{N}{2}\ln\left[SSE(\rho)\right] \tag{19.14}$$

where *SSE* denotes the sum of squared residuals. Since the maximum likelihood estimation of the global SDM model relies on least-squares estimates and the computation of the log-determinant, a recursive spatial estimation method is conceivable. Pace and LeSage (2004, p. 35) develop such a recursive spatial maximum likelihood approach based on recursive matrix decompositions used to compute log-determinants combined with recursive least squares. More specifically, their approach to compute the log-determinant that appears in (19.14) relies on the decomposition of $(I - \rho W)$ into two triangular matrices L and U, that is, $(I - \rho W) = LU$, known as the LU decomposition. It is straightforward to show that:

$$\ln\left|I-\rho W\right| = \ln\left|U\right| = \sum_{j=1}^{N} \ln u_{jj} \tag{19.15}$$

where u_{jj} is the diagonal element in position $(j,j)$of the matrix U. Pace and LeSage (2004) underline the recursive nature of the LU decomposition to design a spatial autoregressive local estimation method for the SDM model. Indeed, the log-determinant of the successive sub-matrices are the successive sums of the logarithms of the diagonal elements of the matrix U, so that we have: u_{11} for the first sub-matrix, $\ln u_{11} + \ln u_{22}$ for the second one, and more generally $\sum_{j=1}^{m} \ln u_{jj}$ for the mth sub-matrix with $m \leq n$.

To implement the estimation procedure for observation i, note that the observations in the sample are first ordered with respect to their distance to observation i. Also, the rows and columns of the weights matrix are consequently reordered. Denote that matrix by W_i. Suppose now that we want to consider sub-samples of size equal to m corresponding to the m-nearest neighbors to observation i. More specifically, the local profile log-likelihood function of Pace and LeSage (2004) is written as follows (omitting the constant term):

$$L_i(\rho_i) = \ln\left|I-\rho_i W_i\right| - \frac{m}{2}\ln[SSE(m,\rho_i)] \tag{19.16}$$

It can therefore be rewritten as:

$$L_i(\rho_i) = \sum_{j=1}^{m} \ln u(\rho_i)_{jj} - \frac{m}{2}\ln[SSE(m,\rho_i)] \quad \textit{where} \quad m \leq n \tag{19.17}$$

The recursive method of Pace and Barry (1997) is then used to estimate ρ_i, which may then be interpreted as the local spatial autocorrelation parameter. We note that as $m \to N$ these estimates approach the global estimates based on all N observations that would arise from the global SDM model. The procedure is then repeated for all the observations in the sample $i = 1,...,N$ yielding a sequence of N spatial autoregressions.

A Bayesian variant of this approach, labeled BSALE, has been developed in Ertur et al. (2007) in the empirical regional convergence framework and applied to a sample of 138 European regions over the period 1980–95. On the one hand, regarding heterogeneity as with the standard GWR approach, the proposed locally linear spatial autoregressive model partitions the cross-sectional sample observations by treating each location along with neighboring locations as a subsample. This avoids arbitrary decisions regarding how to partition the sample observations, but allows for variation in the parameter estimates across all observations. On the other hand, it is assumed that similarities in legal and social institutions as well as culture and language might give rise to local uniformity in economic structures, leading to similar local schemes for convergence speeds and thus to a concept of 'local convergence'. In other words, there should exist spatial clustering in the magnitudes of the β-convergence parameter estimates. However, the locally linear spatial estimation method does not impose a priori similar convergence speeds for spatially neighboring observations. Rather, β-convergence parameters for each region in the sample are estimated based on the subsample of neighboring regions. Furthermore, Bayesian techniques produce robust estimates with regards to potential outliers and heteroskedasticity of unknown form. A Markov chain Monte Carlo (MCMC) estimation method is then developed to implement the proposed approach.

The econometric results obtained using different subsample sizes show clear evidence that indeed the spatially lagged endogenous variable should be included in the specification. As the subsample size increases, they get larger positive modal values for the local spatial autocorrelation coefficients. Individual estimates exhibit local spatial dependence of a sufficiently large magnitude to create bias in standard GWR least-squares estimates even for relatively small subsample sizes. Estimated local spatial autocorrelation coefficients also present a clear country-dependent spatial pattern. Concerning the individual β-convergence parameter estimates, it should be noted that country-level differences are apparent: estimates change abruptly as one moves from one country to another. In addition, there is also substantial variation between regions within a country. Standard statistical inference does not apply here, so samples of draws generated during MCMC sampling are then used to produce confidence intervals. It appears that only 31 regions, mainly located in south-western Europe (Portugal, Spain, some French regions), are converging. All other regions are characterized by non-significant estimates. These conclusions are similar for subsample sizes varying from roughly one-quarter to three-quarters of the sample size. However, it should be noted that the estimates suffer from sample re-use as in the case of other locally linear non-parametric estimation methods preventing interpretation of the results in a strict statistical sense.

One common criticism that can be made to most of the applications of GWR or SALE presented in the growth and convergence literature is the lack of rigorous theoretical foundations, as the estimated regressions are not derived as reduced forms from structural theoretical models embedding both continuous spatial parameter heterogeneity and spatial interaction.[4] To our knowledge, Ertur and Koch (2007) is the first attempt to develop such a theoretical growth model, which leads to the local SDM model as the relevant econometric reduced form to be estimated. More precisely, their augmented Solow model includes both physical capital externalities as suggested by the Frankel–Arrow–Romer model and spatial externalities in knowledge to model technological interdependence. They suppose that technical progress depends on the stock of physical capital per worker, which is complementary with the stock of knowledge in the home country. It also depends on the stock of knowledge in other countries which affects the technical progress of the home country. The intensity of this spillover effect is assumed to be related to some concept of socio-economic or institutional proximity, which is captured by exogenous geographical proximity. Their model provides, as a reduced form, a conditional convergence equation, which is characterized by complete parameter heterogeneity and which is therefore estimated using SALE on a sample of 91 countries over the period 1960–95. Their econometric results support their model as all the coefficients have the predicted signs and underline spatially varying convergence speeds across countries as well as varying coefficients for all other explanatory variables and their spatial lags as the saving rates and population growth rates.

Ertur and Koch (2006b) extend this model by including human capital as a production factor following Mankiw et al. (1992) and propose to model human capital externalities along the lines of Lucas (1988). Technological interdependence is still modeled in the form of spatial externalities in order to take account of the worldwide diffusion of knowledge across borders. The extended model also yields a spatial autoregressive conditional convergence equation including both spatial autocorrelation and parameter heterogeneity as a reduced form. However, in contrast to Mankiw et al. (1992), their results show

that the coefficient of human capital is low and not significant when it is used as a simple production factor. Further research is therefore needed to investigate the role played by human capital in growth and convergence processes. In addition, those models having been developed for countries at the international scale, it would be interesting to figure out what modifications are needed to adapt them at the regional scale to help to better understand regional growth and convergence processes.

19.5 Concluding remarks

This chapter has aimed at presenting various approaches dealing with heterogeneous reaction eventually combined with interaction between neighboring units of observation developed in the spatial econometric literature, in the framework of cross-sectional models, and applied to growth and convergence processes. Discrete and continuous forms of heterogeneity allowing spatial variations in regressions coefficients have been studied. Geographically weighted regressions have been used in the empirical growth and convergence literature to model spatial heterogeneity in regression coefficients, and their generalizations taking into account spatial autocorrelation as well as spatial heterogeneity are especially interesting. Further modeling strategies may include newly developed Bayesian models with spatially varying coefficients (Gelfand et al., 2003) and neural networks (Lebreton, 2005). The toolbox of the applied growth researcher is now very diverse and rich. However, most importantly, we believe that, in further research, more efforts should be oriented towards developing, especially at the regional scale, spatial structural theoretical growth models, which would provide the basis of econometric reduced forms that would be estimated using the spatial econometric toolbox.

Notes

1. See Rey and Janikas (2005) and Rey and Le Gallo (forthcoming) for such a review.
2. See Anselin (1988) for further details on the random coefficients model in a cross-sectional context. See also Brunsdon et al. (1999b) for a comparison between random coefficients models and the GWR model, which is considered in section 19.3 of this chapter.
3. The generalization to more than two regimes is straightforward.
4. Until recently, this criticism was also valid for the simpler spatial specifications of convergence models. Some important contributions by Egger and Pfaffermayr (2006), López-Bazo et al. (2004) and Vayá et al. (2004) fill the gap between theoretical and empirical models.

References

Abreu, M., H.L.F. de Groot and R.J.G.M. Florax (2005), 'Space and growth: a survey of empirical evidence and methods', *Région et Développement*, **21**, 12–43.

Agterberg, F. (1984), 'Trend surface analysis', in G.L. Gaile and C.J. Wilmot (eds), *Spatial Statistics and Models*, Boston, MA: Reidel, pp. 147–71.

Anselin, L. (1988), *Spatial Econometrics: Methods and Models*, Dordrecht: Kluwer Academic Publishers.

Anselin, L. (1990a), 'Spatial dependence and spatial structural instability in applied regression analysis', *Journal of Regional Science*, **30**, 185–207.

Anselin, L. (1990b), 'Some robust approach to testing and estimating in spatial econometrics', *Regional Science and Urban Economics*, **20**, 141–63.

Anselin, L. (1995), 'Local indicators of spatial association: LISA', *Geographical Analysis*, **27**, 93–115.

Anselin, L. (2006), 'Spatial econometrics', in T.C. Mills and K. Patterson (eds), *Palgrave Handbook of Econometrics: Volume 1, Econometric Theory*, Basingstoke: Palgrave Macmillan.

Anselin, L. and A. Bera (1998), 'Spatial dependence in linear regression models with an introduction to spatial econometrics', in A. Ullah and D.E.A. Giles (eds), *Handbook of Applied Economics Statistics*, Berlin: Springer-Verlag, pp. 237–89.

Anselin, L. and W.T. Cho (2002), 'Spatial effects and ecological inference', *Political Analysis*, **10**, 276–97.

Anselin, L. and D.A. Griffith (1988), 'Do spatial effects really matter in regression analysis?', *Papers of the Regional Science Association*, **65**, 11–34.
Anselin, L., J. Le Gallo and H. Jayet (2008), 'Spatial panel econometrics', in L. Matyas and P. Sevestre (eds), *The Econometrics of Panel Data*, Dordrecht: Kluwer Academic Publishers, pp. 625–60.
Basile, R. and B. Gress (2005), 'Semi-parametric spatial auto-covariance models of regional growth in Europe', *Région et Développement*, **21**, 93–118.
Bivand, R.S. and R.J. Brunstad (2005), 'Regional growth in Western Europe: detecting spatial misspecification using the R environment', *Papers in Regional Science*, **85**, 277–97.
Brock, W. and S.N. Durlauf (2001), 'Growth empirics and reality', *World Bank Economic Review*, **15**, 229–72.
Brunsdon, C., A.S. Fotheringham and M. Charlton (1996), 'Geographically weighted regression: a method for exploring spatial nonstationarity', *Geographical Analysis*, **28**, 281–98.
Brunsdon, C., A.S. Fotheringham and M. Charlton (1998), 'Spatial nonstationarity and autoregressive models', *Environment and Planning A*, **30**, 957–73.
Brunsdon, C., A.S. Fotheringham and M. Charlton (1999a), 'Some notes on parametric significance tests for geographically weighted regression', *Journal of Regional Science*, **39**, 497–524.
Brunsdon, C., A.S. Fotheringham and M. Charlton (1999b), 'A comparison of random coefficient modelling and geographically weighted regression for spatially non-stationary regression problems', *Geographical and Environmental Modelling*, **3**, 47–62.
Canova, F. (2004), 'Testing for convergence clubs in income per capita: a predictive density approach', *International Economic Review*, **45**, 49–77.
Casetti, E. (1972), 'Generating models by the expansion method: applications to geographical research', *Geographical Analysis*, **4**, 81–91.
Cleveland, W.S., S.J. Devlin and E. Grosse (1988), 'Regression by local fitting, methods, properties, and computational algorithms', *Journal of Econometrics*, **37**, 87–114.
Cliff, A.D. and J.K. Ord (1981) *Spatial Processes: Models and Applications*, London: Pion.
Dall'erba, S. and J. Le Gallo (2008), 'Regional convergence and the impact of European structural funds over 1989–1999: a spatial econometric analysis', *Papers in Regional Science,* **87**, 219–44.
Desdoigts, A. (1999), 'Patterns of economic development and the formation of clubs', *Journal of Economic Growth*, **4**, 305–30.
Durlauf, S.N. (2001), 'Manifesto for a growth econometrics', *Journal of Econometrics*, **100**, 65–9.
Durlauf, S.N. and P.A. Johnson (1995), 'Multiple regimes and cross-country growth behavior', *Journal of Applied Econometrics*, **10**, 365–84.
Durlauf, S.N., P.A. Johnson and J. Temple (2005), 'Growth empirics', in P. Aghion and S.N. Durlauf (eds), *Handbook of Economic Growth*, Amsterdam: Elsevier, pp. 555–677.
Durlauf, S.N., A. Kourtellos and A. Minkin (2001), 'The local Solow growth model', *European Economic Review*, **45**, 928–40.
Easterly, W. and R. Levine (2001), 'It's not factor accumulations: stylized facts and growth models', *World Bank Economic Review*, **15**, 177–219.
Eckey, H.F., R. Kosfeld and M. Turck (2007), 'Regional convergence in Germany: a geographically weighted regression approach', *Spatial Economic Analysis*, **2**, 45–64.
Egger, P. and M. Pfaffermayr (2006), 'Spatial convergence', *Papers in Regional Science*, **85**, 199–215.
Ertur, C. and W. Koch (2006a), 'Regional disparities in the European Union and the enlargement process: an exploratory spatial data analysis, 1995–2000', *Annals of Regional Science*, **40**, 723–65.
Ertur, C. and W. Koch (2006b), 'Convergence, human capital and international spillovers', LEG Working Paper, no. 2006-03.
Ertur, C. and W. Koch (2007), 'Growth, technological interdependence and spatial externalities: theory and evidence', *Journal of Applied Econometrics*, **22**, 1023–62.
Ertur, C., J. Le Gallo and C. Baumont (2006), 'The European regional convergence process, 1980–1995: do spatial dependence and spatial heterogeneity matter?', *International Regional Science Review*, **29**, 2–34.
Ertur, C., J. Le Gallo and J.P. LeSage (2007), 'Local versus global convergence in Europe: a Bayesian spatial econometric approach', *Review of Regional Studies*, **37**, 82–108.
Fingleton, B. and E. López-Bazo (2006), 'Empirical growth models with spatial effects', *Papers in Regional Science*, **85**, 177–98.
Fischer, M.M. and C. Stirböck (2006), 'Pan-European regional growth and club-convergence: insights from a spatial econometric perspective', *Annals of Regional Science*, **40**, 693–721.
Fotheringham, A.S., C. Brundson and M. Charlton (2000), *Quantitative Geography*, *Perspectives on Spatial Data Analysis*, London: Sage Publications.
Fotheringham, A.S., C. Brundson and M. Charlton (2004), *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, Chichester: Wiley.
Gelfand, A.E., H. Kim, C.F. Sirmans and S. Banerjee (2003), 'Spatial modelling with spatially varying coefficient processes', *Journal of the American Statistical Association*, **98**, 387–96.

Griffith, D.A. (1992), 'A spatially adjusted N-way ANOVA model', *Regional Science and Urban Economics*, **22**, 347–69.
Harvey, A.C. (1976), 'Estimating regression models with multiplicative heteroscedasticity', *Econometrica*, **44**, 461–5.
Le Gallo, J. and S. Dall'erba (2006), 'Evaluating the temporal and spatial heterogeneity of the European convergence process: 1980–1999', *Journal of Regional Science*, **46**, 269–88.
Le Gallo, J. and C. Ertur (2003), 'Exploratory spatial data analysis of the distribution of regional per capita GDP in Europe, 1980–1995', *Papers in Regional Science*, **82**, 175–201.
Lebreton, M. (2005), 'The NCSTAR model as an alternative to the GWR model', *Physica A*, **355**, 77–84.
LeSage, J.P. (2004), 'A family of geographically weighted regression models', in L. Anselin, R.J.G.M. Florax and S.J. Rey (eds), *Advances in Spatial Econometrics: Methodology, Tools and Applications*, Berlin: Springer.
Leung, Y., C. Mei and W. Zhang (2000a), 'Statistical tests for spatial non-stationarity based on the geographically weighted regression model', *Environment and Planning A*, **32**, 9–32.
Leung, Y., C. Mei and W. Zhang (2000b), 'Testing for spatial autocorrelation among the residuals of the geographically weighted regression', *Environment and Planning A*, **32**, 871–90.
Liu, Z. and T. Stengos (1999), 'Non-linearities in cross-country growth regressions: a semi-parametric approach', *Journal of Applied Econometrics*, **14**, 527–38.
López-Bazo, E., E. Vayá and M. Artis (2004), 'Regional externalities and growth: evidence from European regions', *Journal of Regional Science*, **44**, 43–73.
Lucas, R.E. (1988), 'On the mechanics of economic development', *Journal of Monetary Economics*, **22**, 3–42.
Mankiw, N.G., D. Romer and D.N. Weil (1992), 'A contribution to the empirics of economic growth', *Quarterly Journal of Economics*, **107**, 407–37.
McMillen, D.P. (1996), 'One hundred fifty years of land values in Chicago: a nonparametric approach', *Journal of Urban Economics*, **40**, 100–124.
McMillen, D.P. and J.F. McDonald (1997), 'A nonparametric analysis of employment density in a polycentric city', *Journal of Regional Science*, **37**, 591–612.
Mei, C.-L., S.-Y. He and K.-T. Fang (2004), 'A note on the mixed geographically weighted regression model', *Journal of Regional Science*, **44**, 143–57.
Mei, C.-L., N. Wang and W.X. Zhang (2006), 'Testing the importance of the explanatory variables in a mixed geographically weighted regression', *Environment and Planning A*, **38**, 587–98.
Mennis, J. (2006), 'Mapping the results of geographically weighted regression', *The Cartographic Journal*, **43**, 171–9.
Neven, D. and C. Gouyette (1995), 'Regional convergence in the European Community', *Journal of Common Market Studies*, **33**, 47–65.
Ord, J.K. and A. Getis (1995), 'Local spatial autocorrelation statistics: distributional issues and an application', *Geographical Analysis*, **27**, 286–305.
Pace, R.K. and R. Barry (1997), 'Quick computation of regressions with a spatially autoregressive dependent variable', *Geographical Analysis*, **29**, 232–47.
Pace, R.K. and J. LeSage (2004), 'Spatial auroregressive local estimation', in A. Getis, J. Mur and H. Zoller (eds), *Spatial Econometrics and Spatial Statistics*, New York: Palgrave Macmillan.
Páez, A., T. Uchida and K. Miyamoto (2002a), 'A general framework for estimation and inference of geographically weighted regression models: 1. Location-specific kernel bandwith and a test for locational heterogeneity', *Environment and Planning A*, **34**, 733–54.
Páez, A., T. Uchida and K. Miyamoto (2002b), 'A general framework for estimation and inference of geographically weighted regression models: 2. Spatial association and model specification tests', *Environment and Planning A*, **34**, 883–904.
Ramajo, J., M.A. Marquez, G.J.D. Hewings and M.M. Salinas (2008), 'Spatial heterogeneity and interregional spillovers in the European Union: do cohesion policies encourage convergence across regions?', *European Economic Review*, **52**, 551–67.
Rey, S.J. and M.V. Janikas (2005), 'Regional convergence, inequality, and space', *Journal of Economic Geography*, **5**, 155–76.
Rey, S.J. and J. Le Gallo (forthcoming), 'Spatial analysis of economic growth and convergence', in T.C. Mills and K. Patterson (eds), *Palgrave Handbook of Econometrics: Volume 2, Applied Econometrics*, Basingstoke: Palgrave Macmillan.
Temple, J. (1999), 'The new growth evidence', *Journal of Economic Literature*, **37**, 112–56.
Vayá, E., E. López-Bazo, R. Moreno and J. Surinach (2004), 'Growth and externalities across economies, an empirical analysis using spatial econometrics', in L. Anselin, R.J.G.M. Florax and S.J. Rey (eds), *Advances in Spatial Econometrics: Methodology, Tools and Applications*, Berlin: Springer, pp. 433–55.
Wheeler, D.C. and C.A. Calder (2007), 'An assessment of coefficient accuracy in linear regression models with spatially varying coefficients', *Journal of Geographical Systems*, **7**(2), 145–66.

Wheeler, D. and M. Tiefelsdorf (2005), 'Multicollinearity and correlation among local regression coefficients in geographically weighted regression', *Journal of Geographical Systems*, **7**, 161–87.
Yu, D.-L. (2006), 'Spatially varying development mechanisms in the Greater Beijing Area: a geographically weighted regression investigation', *Annals of Regional Science*, **40**, 173–90.

20 CGE modeling in space: a survey

Kieran P. Donaghy[1]

20.1 Introduction: computable general equilibrium models

Quantitative general equilibrium analysis of an economy based on principles of microeconomic reasoning has represented the high ground that theoretical and applied economists have sought to occupy since theories of economy-wide behavior were first formulated. Economists have been interested not only in directions of change in the values of critical variables, induced by policies or exogenous shocks, but also in magnitudes, reflecting compensating systemic adjustments. Economists have also needed the capacity to conduct quantitative general equilibrium analyses to make headway with theoretical arguments in which, because of non-linearities in functional forms or the number and complexity of assumptions, it is difficult to isolate the effects of a change in an assumption. Yet, until relatively recently, because of limitations in data availability or computational know-how and capacity, they have had to settle for quantitative but partial-equilibrium analysis or purely qualitative – and therefore often indeterminate – comparative-statics analysis.

In the last 50 years or so, great strides have been made in what has been termed applied general equilibrium (AGE) or computable general equilibrium (CGE) analysis.[2] By standard accounts – for example Shoven and Whalley (1992) – the emergence of CGE modeling is owed to the happy coincidence of Johansen's (1960) multi-sectoral study of economic growth, Harberger's (1962) analysis of tax policies, and Scarf's (1967) and Scarf and Hansen's (1973) work on conditions for the existence of and algorithms for computation of Walrasian equilibria. Much of the early CGE modeling following on from Scarf and Hansen (1973) was concerned with the effects of taxation and trade policies (Shoven and Whalley, 1984) and analyses conducted with CGE frameworks soon took on a dynamic cast (Pereira and Shoven, 1988; Dixon and Rimmer, 2002). Generalized approaches to specifying, closing, calibrating, solving, simulating and validating CGE models have been developed for different areas of application (see, for example, Shoven and Whalley, 1992; Ginsburgh and Keyser, 1997) and a secondary literature on evaluation of analyses conducted with CGE models has developed (for example Harrison et al., 1993; McKitrick, 1998). Software packages such as GAMS (General Algebraic Modeling System) have had special modules developed to facilitate CGE modeling and such journals as *Economic Systems Research*, the *Journal of Policy Modeling* and *Economic Modelling*, among others, regularly feature studies involving applications or innovations in this line of inquiry.

Over the last 20 years CGE modeling has become a stock in trade of regional economists, agricultural economists, economic geographers, or, broadly speaking, regional scientists as they have carried out both regional and interregional or multi-regional analyses of the various types of policies.[3] (See, for example, Dixon et al., 1982 and reviews of the relevant literature by Kraybill, 1993, and Partridge and Rickman, 1998.) More recently, CGE models have taken on an explicit spatial orientation, as the focus of modeling

exercises has turned to analysis of location-specific impacts of unplanned events and industrial, transportation, environmental, fiscal and other types of policies. Spatial CGE models (SCGEs) have been employed by researchers at various scales of spatial and temporal resolution to forecast and evaluate impacts within a particular local labor market or an area spanning continents ranging from the near term or out to 50 years.

This chapter surveys a broad swathe of studies in the recent literature in SCGE modeling. The articles, book chapters and working papers discussed constitute a (more or less) representative sample of scholarship in this area. The current rate of production and space limitations rule out more exhaustive coverage.

There are at least five discernible streams of literature on SCGE modeling – amounting to over three dozen papers published and unpublished – that can be traced back to the early 1990s.[4] Taking these up roughly in order of appearance, in the first stream of studies are extensions of CGE models developed to study trade issues. The specifications and approaches taken to solve these models followed Johansen (1960), Scarf (1967), Scarf and Hansen (1973) and Shoven and Whalley (1984). Studies of a kindred spirit have extended the multi-regional and interregional applied general equilibrium tradition of Moses (1955) and Isard (1951). A second stream of studies may be viewed as drawing upon two literatures: the spatial price equilibrium literature going back to Enke (1951), Samuelson (1952) and Takayama and Judge (1964) and continued by Batten and Boyce (1987) and Roy (1995), and the predictive freight network literature of Harker (1987) and Harker and Friesz (1986a, 1986b). A third and swelling stream of SCGE scholarship comprises studies influenced by the 'new economic geography' of Fujita et al. (1999). A fourth stream of studies encompasses analyses of environmental policies with spatially extended CGE models on a global scale. And a fifth identifiable stream of SCGE modeling, which can be contrasted with the fourth in terms of range of spatial scope of analysis, concerns development in smaller areas.

The primary objective of this survey is to convey a sense of the scope of SCGE modeling now being undertaken.[5] There have been a number of excellent surveys of studies conducted with CGE models, which have suggested categorial schemes that might be used to compare models and uses to which they have been put.[6] Because SCGE models are being constructed and simulated for such diverse reasons, however, we shall not follow a standard template in discussing the studies surveyed. Rather, we shall focus on what is distinctive about the studies and what the scholars have accomplished by taking the directions they have. The chapter concludes with a discussion of new directions in which SCGE modeling might be taken.

20.2 First extensions of Walrasian CGE trade models into space

Among the first efforts to incorporate a spatial dimension explicitly in Walrasian CGE models developed to examine the effects of trade policies were studies by Jones and Whalley (1989), Buckley (1992) and Wigle (1992). Bröcker (1998a) demonstrated how to operationalize such an SCGE. Isard and Azis sought in their 1998 textbook chapter to lay out what they argue is the preferred approach to specifying such models and critique earlier modeling efforts. In their 1999 working paper, Löfgren and Robinson demonstrated how multi-regional CGE models could be made to accommodate both spatial network models and trade regime shifts. More recently, in their 2005 paper, Ando and Meng presented a multi-regional AGE model for China featuring the transport sector's

behavior. All of these models are static in orientation. Table 20.1 characterizes some of their general features. Because of space limitations, we shall not discuss the papers of Wigle, Bröcker, and Ando and Meng.

Jones and Whalley (1989)

One of the first multi-regional CGE models to be developed to investigate regionally differentiated effects of taxation and international trade policies was that presented and discussed in the 1989 *Journal of Urban Economics* article of Rich Jones and John Whalley. Intended as a flexible tool for evaluating the effects of different policies on the regions of Canada, the model was clearly related to applied general equilibrium models developed earlier by the authors. The model, which comprises six regions of Canada and the rest of the world, is static in orientation. There are both good and factor flows between regions and each region has 13 sectors, which use primary factors – capital and labor services and natural resources – and intermediate products as inputs. (The rest of the world uses only capital and labor services.) On the demand side, demanded quantities are derived from maximization by representative agents of nested CES/LES[7] utility functions subject to budget constraints. In keeping with the so-called 'Armington assumption' made in their earlier work, Jones and Whalley assume that goods are distinguished qualitatively by their location of origin (see Armington, 1969).

There are two scenarios of capital mobility. In the first, capital can move between sectors and regions but not internationally; in the second, capital can move internationally as well. The federal government is a utility-maximizing agent which produces no output (public goods are ignored), but regional governments are not agents. Interregional labor mobility is determined on the basis of comparison of utility between potential locations of residence.

Key federal policies with interregional feedbacks include trade tariffs, transportation subsidies, energy policies, intergovernmental transfers, economic union issues, federal tax system features, regional development programs and agricultural programs. The notion of general equilibrium at work is the Walrasian one of a set of prices at which all markets clear. When capital is internationally mobile, there is only one capital market equilibrium, and when not, there are two. Even though labor is assumed to be homogeneous, endogenous determination of migration results in different wage rates, hence separate labor market equilibrium conditions, which must be satisfied in each market. Model variants allow for changes in elasticity configurations and factor mobility assumptions.[8]

The model is calibrated according to the approach taken with similar models of national economies – model equilibrium conditions and equilibrium data are used to solve for parameter values of the model equations. Two types of equilibria are investigated – benchmark and counterfactual. Exogenously given elasticity estimates allow other parameters to be calibrated. Jones and Whalley use the 1981 'micro-consistent regional data set' for Canada. In this data set, each region is viewed as a separate regional economy. Trade between regions is incorporated, but tax payments from regions to the federal government, intergovernmental transfers received by regions, and federal government purchases of regionally produced goods also appear.[9] International trade elasticities were taken from Stern et al. (1976). It is assumed that interregional trade elasticities are the same as their international counterparts. Energy demand and supply elasticities were taken from various published studies.

Table 20.1 Key features of Walrasian SCGE trade models discussed

SCGE study	Purpose	Dimensions	How space is introduced	Market structure
Jones and Whalley (1989)	Develop a flexible tool to evaluate effects of policies on Canadian regions	6 regions and ROW 13 sectors	Mobile factors, multiple levels of governance, interregional trade	Competitive
Buckley (1992)	Examine in policy scenarios implications of specifications of handling and movement	3 regions and ROW? 5 sectors	Distance-related costs of transport and handling	Competitive
Wigle (1992)*	Examine role transportation costs play in determining welfare effects at national and regional levels	7 regions 13 sectors/commodities	Distance-related costs of transport and handling	Competitive
Bröcker (1998a)*	Demonstrate how to operationalize a prototype SCGE	variable	Distance-related costs of transport and handling and spatial interaction	Competitive
Isard and Azis (1998)	Present ideal specification with theoretical foundations and critique prior studies	NA	Distance-related costs of transport and handling	Competitive and imperfectly competitive
Löfgren and Robinson (1999)	Develop multiregional CGE model capturing regime shift and spatial network	3 regions and ROW 5 sectors/ commodities	Distance-related costs of transport and handling	Competitive
Ando and Meng (2005)*	Provide an SCGE model considering transport sector and price differentials and demonstrate regional characteristics of China	29 regions and ROW 7 industries	Distance-related costs of transport and handling	Competitive

SCGE study	Inputs	Consumption	Production	Factor mobility	How transport costs introduced
Jones and Whalley (1989)	K,L,M,N	CES/LES	CD	K,L mobile	Fixed input coefficients
Buckley (1992)	K,L,M	LES	Multi-level CD/I-O	L mobile K fixed	Method of Round (1988)
Wigle (1992)*	PPF	Nested CES	CET	Factors immobile	Mark-up cost
Bröcker (1998a)*	variable	Nested CES	Nested CES	Factors immobile	Iceberg transport costs Trade pooling
Isard and Azis (1998)	K,L,M,N	AIDS	Nested varieties	All factors mobile	Resource using sector
Löfgren and Robinson (1999)	K,L,M,N	CD	Piecewise linear approximation of CES (combination of Leontief)	Non-agricultural capital is immobile; land and capital are mobile inside agriculture; labor is mobile across sectors but not regions.	Transport network equilibrium pricing
Ando and Meng (2005)*	K,L,M	Nested CD	CD	Factors immobile	Resource using sector

Notes: K, L, M, and N denote capital, labor, intermediate goods and natural resources, respectively. AIDS, CD, CES, CET, and LES denote almost ideal demand system, Cobb–Douglas, constant elasticity of substitution, constant elasticity of transformation, and linear expenditure system, respectively.
* Indicates study not explicitly discussed in this chapter because of space limitations.

In applications of the model, Jones and Whalley compute welfare effects in terms of Hicksian equivalent variations and interregional net labor flows for tax policies, intergovernmental transfers and removal of taxes and subsidies.[10] In so doing, the authors raise the level of quantitative input in the regional policy debate.[11]

They consider the following counterfactual cases:

- Replace federal government tax, subsidy and transfer policies with a yield-preserving uniform rate federal sales tax.
- Replace regional government tax, subsidy and transfer policies with a yield-preserving uniform rate regional sales tax.
- A combination of the first two cases.
- Remove intergovernmental transfers.
- Remove federal interpersonal transfers.
- Remove all federal non-energy taxes.
- Remove all federal energy taxes and subsidies.

In all cases, Jones and Whalley found the regional impacts of policy changes to be significant, and take these findings to suggest that the modeling approach is valuable in studying the regional impacts of changes in fiscal policies at federal and regional levels.

Buckley (1992)

As noted above, many of the early CGE models developed to examine the effects of trade policy did not explicitly account for the costs – in terms of transportation and wholesaling services – of moving goods across space. Patrick Buckley was one of the first to address this omission. Buckley's model most closely resembles Jones and Whalley's (1989) multi-regional model of Canada. There are five sectors, three regions, and four major blocks: production, intra- and interregional trade and clearing-houses, consumption and balance equations. Buckley describes the model's operation as follows:

> In solving the model, fixed quantities of factor inputs are sold to the production block where they are combined with intermediate inputs to create regional products. These products are then sent to clearinghouses using transportation and wholesaling services. Each region has clearing-houses, which receive intra- and interregional flows in addition to imports from the rest of the world. Under an Armington assumption of imperfect substitution . . . the same commodity from numerous origins becomes a regionally specific, clearinghouse-composite good . . . made available for either local consumer demand, producer intermediate input, or international export. At each point where a transfer of mass occurs, payments must flow in the opposite direction. Hence, factor utilization generates income used to pay for consumption and in turn for production and the factor inputs themselves. (p. 333)[12]

In a solution of the model, all markets clear. Labor is intra-regionally mobile, while capital inputs are fixed by sector and region. Two sets of balance equations determine an endogenous set of factor and product prices and create a Walrasian equilibrium. A key difference in this interregional CGE model is that transportation and wholesale services are used explicitly but remain untraded goods.

Buckley uses the model to examine five policy scenarios. It is assumed that policies would lead to changes in production technology, changes in consumption patterns,

changes in factor market constraints, changes in net international trade, and changes in intra- and interregional movement technology and networks.[13]

To explore the implications of the explicit specifications of handling and movement used in his interregional CGE model (ICGE), Buckley compares results obtained with the former with those obtained with an implicit method, along the lines suggested by Round (1988). According to the latter, transportation and wholesale services are treated as commodities that can be directly consumed, traded or substituted interregionally. On the production side, transportation and wholesale services are handled as direct intermediate inputs, rather than origin–destination marginals. On the demand side, consumers and exporters distinguish direct purchases of clearing-house composites of transportation and wholesaling from purchases of tradable products. The implicit approach creates a 'production price' model rather than an explicit 'consumer price' model, which makes it consistent with input–output (I–O) accounting.

Based on comparisons of counterfactual scenario outcomes of an increase in productivity in transportation of one region, Buckley finds falling demand for transportation services in production. The advantages of the ICGE method appear to be threefold (p. 343):

1. 'the direct linkage of movement and handling services to the actual products moved prevents over/under production of those services by regions';
2. 'end users of the products purchase a "correct" amount of services for delivery of their bundle of goods';
3. 'since the explicit method limits consumer final demand to transportation for personal mobility, a much clearer picture of consumption emerges. Thus the explicit method provides a spatially focused method of understanding change in the transportation and wholesaling sectors.'

It has been well appreciated that the solutions of CGE models are very sensitive to their specifications and that the specifications chosen can influence how data are collected and organized. Buckley's analysis has shown that a different pattern of regional production activity emerges from an explicit formulation of the use of resources in movement of commodities across space rather than an implicit one, and consequently conclusions drawn about policies will be affected.

Isard and Azis (1998)

In 1998, Walter Isard and co-authors Iwan Azis, Matthew Drennan, Ronald Miller, Sidney Saltzman, and Erik Thorbecke brought out a new textbook introducing first-year graduate students to the field of regional science. The text differs from the 1960 classic, *Methods of Regional Analysis*, not only in not aiming for completeness in surveying the field, and its selective inclusion of methods – such as spatial econometrics, social accounting matrix analysis, and micro-simulation modeling – developed more recently than 1960, but also in its emphasis on interregional analysis (Isard et al., 1960). In the eighth chapter of the book, Isard and Azis reach the point of subject development where: 'market equilibrium is a basic element, prices are free to vary at least relatively and where they fully impact production, consumption, trade and spatial interaction in general. Both linear and non-linear functions will be involved' (p. 334).

In a step-by-step fashion, the authors of the chapter exposit an approach to applied general interregional equilibrium (AGIE) analysis, which, at the time of the book's publication and to the best of the authors' knowledge, had yet to be implemented, but illustrate fundamental points with numerical examples. A basic concern of the authors, (shared by Buckley, 1992), is that: 'to avoid distortion in understanding interregional economic relations, transportation must be characterized as a resource-using activity which consumes goods and services from other sectors in its own and other regions' (Isard and Azis, 1998). They proceed from the most basic of relationships – within a two-commodity, two-resource, two-region system – treating fundamental issues of trade and location, and then move on to the development of a social accounting matrix to serve as a core for AGIE analysis in which there is consumption, production, transport, government, investment and financial activity in a market system. In so doing, Isard and Azis argue that location analysis can be properly handled within AGIE analysis, in which case there would not be two general theories – one of trade and one of location – but only one. They also argue that some of the prevailing models of trade theory – for example the Heckscher–Ohlin model – can lead to invalid conclusions when travel costs are excluded from consideration.

After setting out their preferred approach, Isard and Azis take up (what they deem to be) problems and questionable assumptions of standard applied general equilibrium (AGE) modeling approaches when such an approach is applied to interregional analysis. With regard to the characterization of consumer behavior, they view as unrealistic specifications that imply constant expenditure shares or a common elasticity of substitution, regardless of the commodities in question. (They recommend using an Almost Ideal Demand System or variant, even though they are aware of the additional data requirements of doing so.) With regard to producer behavior, they find problematic the use of a common production function in characterizing technologies of all industries, and suggest that different specifications might be used for different industries. They also criticize the use of constant elasticity of transformation (CET) functions to allocate production between export and domestic markets because such a practice fails to reflect the interplay of delivered cost pricing and market imperfections at different levels. On the production side, they cite the need to account explicitly for increasing returns and externalities within AGIE models.

Isard and Azis view the introduction of imperfect competition to AGE models as a step forward but are troubled by the lack of realism in the stylized behaviors of actors under monopolistic competition in some models – for example the absence of income effects on perceived demand. An area they identify as needing much development is intertemporal analysis. In particular, they feel that the endogenous determination of expectations is not well handled and they point out that how this issue is resolved will affect such other issues as choice of numeraire, calibration of the base-period equilibrium, and the handling of temporary equilibria. With respect to characterizations of government behavior, Isard and Azis argue that the government should be viewed as a long-run investor and that, when so viewed, feedbacks from associated behavior need to be explicitly incorporated into model specification.

Turning their attention to non-economic factors, Izard and Azis suggest that neoclassically oriented AGIE modelers may have much to learn from structuralist modelers whose approach is characterized by the following features:

- Relevant sets of households and institutions are identified in terms of income flows and possession of wealth.
- Nominal prices and income are used.
- It is recognized that different prices are controlled by different actors.
- Just how much economic rationality is at work in any given situation is questioned – for example agents may act according to rules of thumb.
- An emphasis is placed on capturing key causal relationships.

Isard and Azis conclude their chapter by commenting on what they perceive to be promising work in the area of AGIE analysis. Among the studies discussed in this survey, the authors are approving (although not unequivocally so) of the efforts by Jones and Whalley (1989), Elbers (1996) and Bröcker (1995). They remain critical of the use of Samuelson's notion of 'iceberg' transportation costs, the assumption of pooling of traded goods, the Armington assumption, and the pervasive use of CES functional forms in characterizing firms' technologies and agents' utilities.

Löfgren and Robinson (1999)

Rarely have regionally disaggregated CGE models treated geographical space explicitly or permitted 'regime shifts' for trade flows. In their 1999 discussion paper on spatial networks in multi-region computable general equilibrium models, Hans Löfgren and Sherman Robinson present a model combining the strengths of CGE market simulation models and multi-region programming models. They develop a country-level spatial-network CGE model that is formulated as a mixed-complementarity problem.[14] It is worth revisiting briefly the background of scholarship against which Löfgren and Robinson were working.

At what might be viewed as the dawn of the current era of spatial equilibrium modeling, Enke (1951) and Samuelson (1952) extended the applicability of the 'transportation problem' of the linear programming literature from Hitchcock (1941) and Koopmans (1949) onwards by introducing price-responsive regional demand and supply functions. Samuelson's formulation showed that:

> the problem of maximizing 'net social payoff' (the sum of consumers' and producers' surpluses in the different regions less transportation costs) subject to regional commodity balance equations generates a set of optimality conditions that define equilibrium in each regional market. (Löfgren and Robinson, 1999, p. 3)

The contribution of Takayama and Judge (1964) was to generalize Samuelson's approach to multiple products and show that, if the conditions for solution to Samuelson's problem were written in terms of linear supply and demand functions, the resulting model could be solved with available quadratic-programming algorithms.[15]

The partial-equilibrium models of Takayama and Judge (1971) and others maintained the basic trade treatment of the original transportation problem. In the economy-wide tradition, Isard (1951) developed a methodological approach to specifying and solving a multi-regional – or as Isard prefers, inter-regional – input–output model. Isard's model comprised a set of linear equations with fixed production coefficients (which exclude input substitutability), no supply constraints (hence no price adjustments), and fixed trade coefficients.[16]

To work around Isard's fixed trade coefficients, modelers have employed either an Armington assumption or a trade pool device (both of which have been discussed above). Löfgren and Robinson contend that models that make use of such devices still do not allow for endogenous regime shifts in trade structure. Moreover, while such models may include a transportation structure, rarely do they explicitly account for space.

In contrast with multi-regional (or interregional) I–O models, multi-regional CGE models tend to be characterized by endogenous price determination, price-responsive input substitution and constrained factor supplies. However, such models characterize interregional trade as a spatial network without regime shifts.

Löfgren and Robinson intend to develop a model that can capture both a regime shift in trade structure and the workings of a spatial network. In the interest of simplicity, they do not treat savings or investment at all, and treat government behavior minimally.[17] The country whose regional economies they seek to model (a stylized version of Mozambique) is divided into separate domestic regions represented as connected points in space. The regional economies are characterized by representative households, factors of production, and commodity-producing activities. The technology is Leontief, and Löfgren and Robinson employ a piecewise linear approximation of a neoclassical production function to effect factor substitution while allowing for regime change (which occurs when an activity goes to zero from a positive level or vice versa). Production levels of commodities are determined in regional markets assumed to be perfectly competitive. Transportation services are also provided by a fixed-coefficient technology and prices are endogenously determined. Commodities are perfect substitutes – that is, not differentiated according to region of production or use. The country is a price-taker on international markets and the price of any good imported is assumed to exceed its export price. There can be endogenous regime shifts in trade, production and exchange in factor markets.

The spatial make-up of the economy is as follows. There are two rural regions, one urban region with a port, and a border region. Except for the border region, each region has a representative household, up to three factors of production, and up to five commodity-producing activities. The urban region has no agricultural development.

The baseline solution of the model is calibrated to replicate a social accounting matrix (SAM). Löfgren and Robinson use the model to simulate the effects on the country's economy of a rise in the price of a high-volume export crop and reduced interregional transportation costs. Results obtained demonstrate the ability of the model to capture threshold effects and a diverse pattern of regional impacts. The results point to potentially important complementarities between improved penetration of export markets and investments in domestic transportation networks. Löfgren and Robinson's model permits analysis of issues of structural change, which are intended effects of governmental and non-governmental programs, and thereby meets some of the concerns voiced by Isard and Azis (1998).

20.3 SCGE models in the tradition of spatial price equilibrium and predictive freight network modeling

The next two studies to be discussed confront the problem of specifying and solving a model determining quantities of goods to be produced and shipped, their mill and delivered prices, and the paths by which goods are to be conveyed from their origins to their destinations. The authors of the studies adopt different approaches to solving the

problem – making use of simplifying assumptions, exploiting a division of labor between different modules, and proposing a problem formulation that can exploit solution properties of variational inequality problems. Other studies not discussed here, because of space limitations, but which adopt other alternatives to confronting this problem (and in the interest of evaluating potential impacts of transportation infrastructure investments) are Roson (1996) and Kim et al. (2004).

Elbers (1996)

In his 1996 chapter on linking regional CGE models, which expands on the methodology employed in his 1992 dissertation, Chris Elbers observes that (as of the early 1990s), CGE modelers and spatial price equilibrium (SPE) modelers have differed over whether or not commodities produced by the same sectors in different locations should be treated as homogeneous or as heterogeneous (distinguished by location of origin), with CGE modelers (along with Armington) usually assuming the latter and SPE modelers the former. Another difference is that SPE models are in some respects partial-equilibrium in orientation in that they ignore the fact that transport services are produced from inputs which are produced in other sectors of the economy. But, he notes, accounting for the production of transportation services from model inputs complicates the computation of global general equilibria considerably. Elbers therefore proposes a hybrid model for linking CGE models of different regions that incorporates the Armington assumption for some sectors but incorporates core assumptions of SPE theory for others.

Elbers remarks that most CGE modelers tacitly assume that a productive sector's output consists of only a single homogeneous good and can be represented by a single production function characterizing the technology of a single representative firm. This approach, however, is untenable in the case of transportation because the transport industry comprises multiple 'firms' and 'products' and serves multiple origin–destination pairs. In spite of much trying, Elbers acknowledges that he cannot offer a satisfactory set of assumptions that will result in a multi-regional CGE model with endogenous production of transport services and which is still easy to solve.

The approach that he takes (and employed in his 1996 study of Nepal) is to assume that if a path minimizes transport costs from one region to another, then it minimizes the use of every input used in transportation between the regions as well. If this assumption holds, then link flows – and from them, input demands of the transport sector – follow (generically) uniquely from a spatial equilibrium. Point-valued input demands by the transport sector in turn enable a multi-regional equilibrium solution to be obtained by iterating on a sequence of regional equilibrium computations. By keeping the sub-model of the transport sector relatively simple and employing existing algorithms for solving spatial price equilibrium problems, it becomes possible to include spatially homogeneous goods in a multi-regional CGE model.

Elbers has reservations about the relatively primitive state of the characterization of the transport services-producing sector and the strong assumption of competitive market clearing in the transportation sector.

Friesz, Suo and Westin (1998)

Perhaps the study that goes the furthest in sketching an approach to a general solution to the problem of integrating CGE and SPE models is the 1998 chapter on freight network

and CGE models by Terry Friesz, Zhong-Gui Suo and Lars Westin (hereafter FSW). In arguing for the importance of this work, the authors state:

> A highly accurate inter-regional, inter-modal freight network forecasting tool which employs a detailed representation of the actual freight transportation network . . . is critical to federal policy and decision making related to regulation/deregulation, for it allows volumes, costs, modal splits and the like to be estimated. It also allows the region-specific impacts of transportation policies to be determined. (p. 212)

A principal point of their chapter is to elaborate a methodology that allows the transportation sector to be 'represented in a detailed and theoretically precise manner within a general equilibrium model of [an entire national] economy' (p. 212). In discussing the shortcomings of predictive freight network models, the authors observe that large-scale models of this ilk have not been integrated with computable general equilibrium models for the purpose of forecasting consistent national or regional economic activity levels and prices on the one hand, and detailed freight flows on the other.[18] In response to this particular shortcoming, their study proposes a spatial computable general equilibrium model.

Friesz, Suo and Westin (FSW) observe that a generalized SPE problem derives transportation demand from production and consumption characteristics of spatially dispersed markets and that a generalized SPE model can substitute for a Wardropian shipper's model.[19] If trip generation is introduced through an SPE sub-model, an SCGE model with a detailed representation of the transportation sector can be developed. The authors note that such a model is of limited practical significance because supply and demand functions are required for each commodity and spatially distinct market.[20]

FSW depart from a general formal statement of a perfectly competitive CGE model and show that the solution to this model has the form of a non-linear complementarity problem.[21] They then show that the case of a Leontief technology corresponds to the case of a constant activity analysis matrix – mapping resource use into activities at different locations – and that the Leontief assumption allows the original problem to be cast in the form of a variational inequality problem (VIP).[22]

FSW then turn to the question of how to integrate a detailed freight network model with the solution conditions of the original perfectly competitive CGE model and demonstrate how transportation origin–destination flows can be calculated *ex post* if the original complementarity problem has been solved for a particular set of regions.

Having established how to determine origin–destination flows, FSW next need to adopt a routing principle to allocate traffic flows through a transportation network. For expository purposes they work with a simple Wardropian user equilibrium.[23] They show that under the weak assumption of positive transportation costs, the conditions a Wardropian user equilibrium must satisfy can be put into the form of a non-linear complementarity problem.

FSW combine the original complementarity problem with the Wardropian user equilibrium problem (in complementarity problem form) and state that the vector of prices, activity levels, commodity path flows and minimum transport costs solves the spatial computable general equilibrium model – which is a function of resource endowments, commodity demands, input–output relations and transport costs – if and only if the equation system defining the original problem (using regional variables and parameters) and the system defining the second problem are satisfied together.

As noted above, complementarity problems also lend themselves to formulation and solution as VIPs. VIP formulations of the SCGE problem are possible, allowing one to use results to establish existence and uniqueness of solutions and efficiently compute them. Such formulations allow the simultaneous and consistent calculation of economic activities and detailed transportation network flows. But there are difficulties and disadvantages that inhere to the use of such approaches, which FSW discuss. Problems also arise with formulation and solution of cooperative and bargaining games participated in by carriers, because of non-convexities which strategic behavior introduces.

FSW identify for future research the investigation of cost mark-up strategies for setting rates, constraints to control the extent of spatial arbitrage, and the nature and effects of constraints which endogenously aggregate shippers' delays and costs from carriers' delays and costs.

The authors also question the appropriateness of a static equilibrium concept in characterizing the systems behavior being modeled. They suggest that a dynamic spatial computable general disequilibrium framing of the activities in question may be more suitable.

20.4 SCGE models with new economic geography foundations

The so-called 'new economic geography' (NEG) associated principally with Fujita, Krugman and Venables (1999), which purports to explain where economic activity occurs and why, represents an integration of economic theories in industrial organization, international trade and economic growth. The approach taken by these and like-minded authors emphasizes interaction between increasing returns to scale, transport costs, mobility of productive factors, and backward and forward linkages in a process of cumulative causation. In their 1999 book, *The Spatial Economy*, which is in effect an edited collection of papers published by the authors over the previous decade, Fujita, Krugman and Venables (FKV) do not engage in applied analysis, per se. But in a concluding chapter, in which they discuss 'the way forward', they identify four directions for future work: enlarging the theoretical 'menu', buttressing the theory with empirical work, going from hypothetical calculation to real quantification, and addressing the welfare and policy implication of the whole approach.

With respect to the first of these directions, they state they 'believe it would be useful . . . to inquire into the behavior of models in which multiple centripetal and centrifugal forces are operating, to ask how the predictions of these models depend on the relative importance of these forces' (p. 346). Such inquiry would enable the findings of empirical research to be interpreted. Empirical work clearly tied to the theoretical models is needed to sort out which of the possibilities suggested by the models are relevant and where further elaboration of the models is necessary. By quantification, FKV mean a 'theoretically consistent model whose parameters are based on some mix of data and assumptions, so that realistic simulation exercises can be carried out'. Certainly, spatial CGE models are of this type. While FKV would like to be able to develop 'computable geographical equilibrium' models, they acknowledge that such modeling is not easy.[24] They suggest that new technical tricks may need to be introduced to make models consistent with data and they hypothesize that 'the payoff to such modeling would . . . be a major step toward making theoretical economic geography an actual predictive discipline, able to evaluate the impacts of hypothetical shocks – including policy changes – on the economy's spatial structure' (p. 348). In *The Spatial Economy*, FKV avoided discussing welfare implications

of structural changes or policies, although such implications have been the focus of Baldwin et al. (2003).

In a number of studies with SCGEs, both anticipating and following the publication of *The Spatial Economy*, various authors have used precepts of the NEG to facilitate analysis of impacts of various developments – for example transportation infrastructure investments or policies to mitigate urban sprawl – or used an SCGE framework to investigate what roles various assumptions of the NEG play in model outcomes. Virtually all of the NEG SCGE models employ a Dixit–Stiglitz (1977) formalization of consumption and production, assume that there are increasing returns to scale in at least some sectors, in which there is also imperfect competition, and assume that all goods are purchased by all sectors and representative households and consumers in all locations. They tend to differ along the lines of:

- the purpose of the model's construction and operation;
- how space is introduced;
- the number and types of sectors and types of actors;
- the formalization of transportation costs;
- how the model is solved or closed and what, then, constitutes an equilibrium solution;
- whether or not the model is static or dynamic in orientation;
- how the model's parameters are calibrated; and
- what the important findings of the model construction or simulation exercise were.

Key features of the NEG SCGE models including those discussed in this section are summarized in Table 20.2. A discussion of important representative examples of SCGE models influenced by aspects of the NEG follows.

Bröcker (1995)

In 1995, Johanes Bröcker published a paper demonstrating how one might go about formalizing much of the NEG in an SCGE, and especially how to introduce Chamberlinian monopolistic competition. Bröcker remarks that, while short on detail and sound empirical foundations, the NEG models in the tradition of FKV have a sophisticated theoretical foundation and complex non-linear specification, which permits them to characterize economies of scale, external economies of spatial clusters of activity and continuous substitution between productive factors in the case of firms, and between different consumer goods in the case of households. Moreover, monopolistic competition of the Dixit–Stiglitz type allows for heterogeneous products, implying variety, and therefore allows for cross-hauling of close substitutes between regions. Bröcker's purpose in publishing his 1995 paper was to demonstrate how 'input–output analysis, gravity modeling, the theory of the intra-industry trade and the theory of general equilibrium under conditions of monopolistic competition could be integrated in a common, logically consistent framework' (p. 148). In addition to possessing a closed-system, market clearing general equilibrium solution of the monopolistic-competition type, his model could be calibrated and solved, was multi-sectoral and multi-regional.

Bröcker observed that the tricks allowing for a tractable model of monopolistic equilibrium with product diversity are due to Dixit and Stiglitz (1977), and have been

Table 20.2 Key features of NEG-based SCGE models discussed

Study	Purpose	Dimensions	Mobility	Nature of transportation costs	Equilibrium solution / closure	Major findings
Bröcker (1995)	Prove feasibility of modeling approach	NA	?	Iceberg	Equality of wage rates	Approach is feasible
Kilkenny (1998)	Examine point of theory	2 regions 2 types of HH Number of firms endogenous	Firms and workers mobile	Uniform delivered price	Equality of wage rates	Relationship between relative transport costs and urban concentration is non-monotonic
Fan et al. (2000)	Proof of concept, examine point of theory	8 cities; 55 cities	Firms and workers mobile	Iceberg	All markets clear, utility is common in all locations	Lower transport costs lead to monocentric core–periphery pattern
Nordman (1998)*	Compare approaches to measuring welfare	3 regions 3 sectors	Factors are immobile	Constant mark-up	Utility is common in all locations	In presence of IRS, benefits from infrastructure must be measured on an economy-wide basis
Venables and Gasiorek (1999)	Impact analysis of infrastructure projects	Variable by country	Depends on simulation	Linear function of distance	Equilibrium concept/closure rules vary by simulation	Spillover effects of projects can vary considerably
Knaap and Oosterhaven (2002)*	Impact analysis of infrastructure projects	548 communities 14 sectors	Workers immobile, firms mobile	Modified iceberg	Labor market, quantity-oriented and demand-constrained	Largest project would shift 8000 jobs from West to North NL
Freidl et al. (2006)*	Demonstrate functional linkage	2 regions 1 sector	Workers mobile	Passenger costs of commuting	Supply equals demand in labor and housing markets	Appropriate response to sprawl includes development of public transport and use of cordon pricing

Note: * Indicates study not explicitly discussed in this chapter because of space limitations.

extensively applied to intra-industry trade modeling (Ethier, 1982), to endogenous-growth theory (Grossman and Helpman, 1991) and, more recently, also to spatial economics in the work of Krugman (1991).[25]

In Bröcker's model the spatial dimension is introduced through a version of Samuelson's (1954) iceberg transportation technology. Analogous to firms, households prefer variety with respect to the brands of goods produced by firms. Deriving the conditions defining a general equilibrium for a closed economy and stepping through the calibration procedure, Bröcker demonstrates that: 'it is possible to calibrate a benchmark equilibrium with a data set that contains no more information than that required for the standard perfect competition approach' (p. 148). Owing to the non-linearity inherent to the model, uniqueness of solution cannot be guaranteed in counterfactual comparative-statics analysis. But, Bröcker argues, finding a unique equilibrium path for small parameter changes is likely.

While Bröcker's paper only demonstrated that Chamberlinian monopolistic competition and other features of the NEG could be implemented in principle in an SCGE model, Bröcker argued that such a model should be appealing to applied modelers for several reasons. Monopolistic competition is more realistic than perfect competition, it frees modelers from having to make the Armington assumption and introducing associated ad hoc parameters, using the theory of intra-industry trade to account for the choice of a supply region; the model offers a natural micro-foundations explanation of why trade flows should obey the gravity equation and it introduces a forward linkage effect that works through market size. Bröcker called for studies comparing results of perfect and Chamberlinian monopolistic competition.[26]

Kilkenny (1998)

In the literature of spatial economics, the effects of reductions in costs of transportation and communication on the economies of rural areas are ambiguous. Such reductions have been hypothesized to promote greater connectedness with urban areas but also promote concentration of activities in urban areas to the detriment of rural areas. These developments are also seen to be accompanied by welfare losses. It is possible, however, that the effects on rural areas of reductions in costs of transporting industrial goods that are initially negative may be ultimately positive, depending on how the ratio of agricultural transport costs to industrial transport costs is affected. To make headway in untangling the relationships involved, quantitative analysis is necessary.

In her 1998 study of transport costs and rural development, Maureen Kilkenny constructs and simulates a two-region (urban and rural) SCGE to demonstrate how transport cost reductions affect rural economic diversity and population. The model is characterized by costly industrial and agricultural transport, both technical and pecuniary economies of scale, product differentiation and uniform delivered price, and a general equilibrium orientation with both firm and worker mobility. The model's specification is intentionally kept as simple as possible to maximize transparency.

In the two regions land may be used for farming and industrial production; the rural region has the larger proportion of farmland. There are two types of households: farmers and industrial workers. Farmers are immobile, but workers can migrate from one region to the other. Initially, all firms and all industrial workers are located in the industrial region. Following in the NEG tradition, monopolistic competition is assumed. Each firm produces a differentiated product, each consumer demands a positive quantity of every variety. Firms

and farmers incur delivery costs, which are constant and linear. Under the assumption of Dixit–Stiglitz consumer preferences and production technologies, the degree of product differentiation is represented by the parametric elasticity of substitution characterizing consumer preferences. Consumers manifest a love for variety and firms enjoy economies of scale as the number of intermediate inputs increases. In a fully employed budget-constrained general equilibrium, consumers' budget share spent on manufactures is the proportion of the total national population in the workforce. An important property of the general equilibrium solution is that agricultural transportation costs raise average fixed and variable costs of industrial production through the wages paid to workers.

A distinctive feature of Kilkenny's model is the assumption of uniform delivered pricing, which she takes to be more realistic than the typical NEG assumption of constant mill pricing and 'iceberg' transportation costs. Under this assumption profits vary across locations. The profit-maximizing delivered price to local markets is a parametric mark-up over the local wage. Within each region, each industrial firm charges the same delivered price to local residents. The activity level per firm is a function of the population size and income distribution, scale and taste parameters. The number of individual firms, hence varieties produced and consumed, that can exist in each region is limited by the regional labor force. Kilkenny depicts decisions of workers to migrate and of firms to relocate as reflecting threshold effects that may be captured in the model's specification through complementary slackness conditions.

The mathematical model consists of first-order necessary conditions for a solution and complementary slackness conditions that allow for corner solutions. Employing a social accounting matrix table, Kilkenny demonstrates that the numerical solution of the spatially diversified model economy satisfies equilibrium conditions in the goods and factor markets as well as regional balance of payments.

Kilkenny employs the model in testing, in a non-statistical sense, hypotheses of Krugman (1991) and Nerlove and Sadka (1991). First she considers behavior of the model under Krugman's assumptions that manufacturers charge the same mill price regardless of location and consumers allocate expenditure over regional good aggregates, when transport costs for agricultural and industrial goods decline at the same rate, by simulating the model with transport cost reductions under constant mill pricing. Kilkenny demonstrates that the qualitative conclusions Krugman arrived at via comparative statics analysis are borne out numerically and, moreover, that symmetric transportation cost reductions favor urban concentration. She also finds numerical support for Nerlove and Sadka's qualitative finding that declining agricultural transport costs, *ceteris paribus*, promotes movement to the city. Under Kilkenny's assumption of uniform delivered price, however, when reductions in transport costs are not symmetric and costs of transporting industrial goods decline more rapidly than costs of transporting agricultural goods, a different result emerges. The relationship between transport costs and urban concentration is non-monotonic, with rural areas first declining and then increasing in diversity and population. Another important finding of Kilkenny's analysis is that in such an outcome the welfare of all types of actors is improved.

Fan, Treyz and Treyz (2000)

Wei Fan and Frederick and George Treyz (FTT) have responded to FKV's call for SCGE studies – and with an increasing number of centripetal and centrifugal forces – with a

conceptual model based on NEG assumptions of increasing returns to scale and monopolistic competition, but also land use, labor mobility, inter-industry purchases and multiple locations in one- and two-dimensional space. The research is in the spirit of 'proof of concept' and examination of the roles played in outcomes by key assumptions.[27]

The assumptions of the model are as follows. Workers are mobile between sectors and regions, as is capital. Land use factors explicitly in consumption and production for all sectors. Differentiated inputs are used in production and geographic space is discrete, so that numerical solution methods may be applied. The agglomerative forces are the price effects and wage effects. The dispersion force is demand for a limited supply of land. Transportation costs take the iceberg form.

The model is solved by employing the dynamic evolutionary algorithm introduced by Fujita and Mori (1997). Given an initial population distribution and autarkic equilibrium with a particular industrial composition, the economy evolves according to a set of laws of motion until, in full equilibrium, all markets clear and utility levels are the same in all regions. FTT observe that: 'the key to the evolutionary methods is that the industrial composition in each region is determined endogenously' (p. 681). As in the case of less complex models of this ilk, the equilibrium solution is path-dependent. With the model's non-linearity contributing to the possible existence of multiple equilibria, 'the economy's off-equilibrium behavior matters a great deal to the final equilibrium' (p. 681).

FTT conducted 1000 simulations with an eight-city model, starting from random initial conditions and leading to path-dependent equilibrium outcomes. They found that while there were cities that were solely agricultural and solely service-oriented, and cities that were agricultural and manufacturing-oriented, and manufacturing and service-oriented, there were no cities that were agricultural and service-oriented (because forward and backward linkages were weak). FTT conducted sensitivity analyses with respect to adjustment speeds and found that differences in adjustment rates – say, in migration – can lead to different, occasionally asymmetric distributions. They considered urban systems of 55 cities. They found that reductions in transportation costs changed equilibrium configurations considerably. As in Kilkenny (1998), lower transportation costs led to more uniform metropolitan population densities. When transportation costs were reduced to one-tenth of the original level, the new equilibrium configuration closely approximated a monocentric core–periphery structure.[28]

Venables and Gasiorek (1999)

One of the most important contributions to the development of the literature of the NEG has been Anthony Venables's 1996 paper on the equilibrium locations of vertically linked industrial industries. In a series of studies for the European Commission, Venables and Michael Gasiorek (1996, 1999) have employed an SCGE model based on the theoretical framework developed in Venables (1996) to investigate the long-run supply-side effects of 'cohesion fund' projects in Europe.[29] (Only the report from the 1999 study is discussed here.) In particular, they seek answers to questions about how changes in regional transportation infrastructure brought about by cohesion fund projects affect the attractiveness of regions as locations for economic activity, hence the location decisions of industries, and their impacts on wage levels and income in different regions and in aggregate. Venables and Gasiorek see the primary innovation of their approach to be capturing the linkages between the actions of firms. The microeconomic detail of their SCGE model

allows them to articulate and trace impacts over network structures that an econometric model would not. (Of course, the disadvantage of the approach is that the model is not subject to empirical modeling or testing.) The model is, in several instances, fitted to data as accurately as possible, but the behavioral assumptions and specification are imposed without testing. Their analysis is intended to complement more traditional cost–benefit analysis of infrastructure projects. The point of any general equilibrium analysis is to capture as fully as possible the effects of a project that occur through the adjustments made throughout an economy that elude partial-equilibrium analysis.

The effects that Venables and Gasiorek attempt to break out are the direct effects that would occur at unchanged levels of activity in the economy, the induced effects arising from changes in activity levels, and spillover effects arising from the spreading of effects through the economy. They believe the wider scope of their analysis is warranted because of the increasing importance of externalities and increasing returns to scale (IRS) in economic analyses and the possible operation of mechanisms of cumulative causation. Venables and Gasiorek argue that to capture cumulative causation adequately in a regional setting a model must have IRS and an imperfectly competitive market structure in a multi-regional general equilibrium setting. They observe that 'such models are inherently complex, and are likely to give rise to multiple equilibria and complex dynamics' (Venables and Gasiorek, 1999, p. 11). Also important in such a model, in their view, is an explicit depiction of input–output relations between firms, enabling backward and forward linkages to be traced (cf. Hirschman, 1958).

Venables and Gasiorek claim that theirs is the first study to attempt to apply an NEG SCGE to a large data set and use it for the purpose of policy analysis, hence we allocate more space to its discussion. They do not characterize technical externalities, but the combination of IRS at the level of the firm, imperfect competition in some but not all sectors, and input–output linkages give rise to potential pecuniary externalities and spillovers.

In their study, the same basic model structure is applied to the economy of each of the cohesion countries – Spain and Portugal as an Iberian amalgam and Greece and Ireland – at varying levels of industrial and regional disaggregation. Regions are taken to be points in space linked by a transport network. There is a high level of spatial disaggregation – for example in the combined model of Spain and Portugal there are 22 regions. Each region has an endowment of primary factors – skilled and unskilled labor and capital. Depending on the simulation experiments conducted, different assumptions about labor and capital mobility apply. External trading partners are assumed to be the rest of the European Union and the rest of the world. Industrial sectors – 16 in Spain and Portugal and fewer in Greece and Ireland – are characterized by either perfectly or imperfectly competitive markets. In perfectly competitive industries, a single homogeneous product is produced in each industry in each region, although differentiation by region of origin is allowed for (the Armington assumption). Each imperfectly competitive industry operates in each region, and each region contains some number of firms from each industry. Venables and Gasiorek use the standard Dixit–Stiglitz representation of IRS and imperfect competition, which involves certain assumptions on the demand system and the assumption that each firm is constrained to operate in a single region only, produces a distinct variety of product and sets price at a constant mark-up over marginal cost. Varieties are assumed to be symmetrical – that is, they are produced with the same technology and face the same demand functions. Hence, if they are sold at the same price, they will enjoy the same level of sales. All

sectors use as inputs the primary factors and intermediate goods. Each sector's output is used as a final and intermediate good. As a consequence of product differentiation and the symmetry assumption, all firms at a particular location in a particular industry supply final demand and intermediate demand in the same proportions.

Venables and Gasiorek consider cases where the number of firms is held constant (in the short run) and allowed to vary (in the long run). In the long-run equilibrium the location of firms is determined by the 'zero abnormal profits' condition, which holds for all industries and regions. The authors want to capture the possibility of intra-industry trade in each industry between all locations. The volumes of interlocational flows are regulated by transport costs between locations. Since data on interregional trade flows are lacking, flows are inferred.

Some regions serve as ports and the imports that enter the country through them are available at fixed world prices. Changes in costs of bringing goods in through a port will lead to substitution of port use. Each port also serves as a source of demand for domestically produced goods. The economy under study is a price-taker with respect to imports but not with respect to exports.

The dimensions of the model are high (in the case of Spain and Portugal, there are 22 x 22 trade flows). The authors achieve tractability by making use of aggregation properties consistent with the Dixit–Stiglitz formulation of product differentiation. They point out that it is possible to construct price indices for each industry at each location. These indices summarize in an exact way the prices of all varieties of product supplied and serve as the basis for consumer choice and firms' import decisions.

Venables and Gasiorek claim that their modeling framework captures the intended effects – for example a decline in transport costs along some link leads to declining prices of products in different regions, affecting firm sales and profits, the entry and exit of firms. The entry of a firm has four discernible effects in this model: it increases competition, reducing profits in the industry; it bids up factor prices, reducing profits; it increases demand for other firms' sales as intermediate products (the backwards linkage) and it increases profits of supplier industries in the location of firm entry (the forward linkage); and, insofar as the new firm supplies intermediate products, it reduces the price index of these products in the location of firm entry.[30]

Venables and Gasiorek had the use of regional data on sectoral output and employment at the NUTS II level of disaggregation for Spain and Portugal, at a level reconcilable with NUTS II for Greece, and at the NUTS III level for Ireland.[31] They were able to obtain from national I–O tables sectoral shares in output for Spain and Portugal but not Greece and Ireland. So they used Spanish coefficients as proxies and reaggregated Greek and Irish data. It is not clear from their write-up how elasticities were calculated. There are no data on interregional trade in Europe. However, using data on trade costs and a distance matrix, Venables and Gasiorek adjusted the relationship between distance and transport costs until the model could reasonably replicate data on aggregate trade volumes. They then 'allowed the relationship between distance and trade volumes to vary across industries (while preserving aggregate trade flows) in order to capture different transportation costs in different industries' (p. 19). They also employed external trade and production data to measure tradability of output of different industries.

The cohesion projects considered by Venables and Gasiorek included the North–South road link in Ireland, the Madrid ring road and Rias-Bajas motorway in Spain, the Tagas

crossing in Portugal, and the Egnatia motorway and Pathe motorway in Greece. The road projects are seen as reducing the effective distance between network nodes. How particular routes would be affected by particular projects was conjectured by the investigators. The effects of projects were assessed in four ways:

1. Direct effects were computed, holding trade flows constant.
2. Effects were computed holding locations of firms and workers constant, but output and sales were allowed to change.
3. The number of firms operating in each industry and region were allowed to change.
4. Workers were allowed to move in response to interregional wage differentials.

Venables and Gasiorek offer the following assessment of what the simulations demonstrated:

> These simulated experiments demonstrate the difference in spillover effects resulting from infrastructure projects of very different natures. They highlight how one project located in a single region . . . can create large welfare benefits rippling through numerous regions, while another project also located in a single region . . . can have much more limited regional spillover effects. Furthermore, they offer explanations for how improved transport networks can lead to gains in welfare for all regions, but can inversely affect the labour income of some regions and positively affect the labour income of others. Finally, the experiments give particularly useful insights into the effects of extremely large projects . . . illustrating the possibility of positive interactions between infrastructure projects. (Venables and Gasiorek, 1999, pp. 58–9.)

Because the specification of the model has been assumed to be correct and is not subject to rigorous verification, Venables and Gasiorek subjected it to extensive sensitivity analysis and found the main results to be quite robust. They suggest that to exploit fully the potential of the modeling approach, other modelers should ensure that good descriptions of projects to be evaluated can be obtained and the projects to be analyzed are large or constitute a series of projects. Their approach provides a 'big picture' of regions as a whole and the economic interaction between regions.

20.5 Analysis of environmental policies in spatial CGE models

Increasingly, CGE models are being employed in studies of potential economic impacts of climate change or mitigative and adaptive responses to climate change. Li and Rose (1995) is representative of studies conducted at the state (regional) level, whereas Nordhaus and Yang (1996) is representative of studies conducted at the multi-regional national level. The next two studies to be discussed use the spatially extended GTAP-E model to consider multinational implications of climate change and policy responses.[32]

Wang and Nijkamp (2007)

Success in mitigating and adapting to global climate change will entail close cooperation between countries with developed and developing economies. Such cooperation is more likely if it can be demonstrated to benefit both types of countries. The notion of a 'clean development mechanism' (CDM) is that of an instrument designed to facilitate the cooperation between countries which are sources of development aid and receiving countries, enabling the former partly to meet reduction targets and the latter to increase their level

of clean technology. Shunli Wang and Peter Nijkamp aim to model and evaluate the implementation of CDMs in a general spatial equilibrium framework by deploying an adjusted version of the GTAP-E model, which incorporates a fine-grained specification of energy sources and technology (Burniaux and Truong, 2002).[33] GTAP-E is an applied general equilibrium model with a data base for the world economy and ecological systems. It can be characterized as a spatially extended CGE model for analyzing the interaction between ecological and economic systems.

As in most neoclassically based CGE modeling frameworks, in GTAP-E overall utility is maximized by decomposing the economic decision process into two sub-problems. Producers maximize profits subject to technological feasibilities, while consumers maximize utility subject to their budget constraint. Price-taking behavior is assumed on both sides of the market. In this model there are three aggregate goods: consumer goods, government expenditure and savings. There is a representative consumer for each region of the world who earns income from natural resources, capital and labor. On the production side, energy–capital substitution is important. The energy–capital composite has a CES form and there is a nested aggregate of energy. GTAP-E can trace back carbon emissions to six categories of sources in each region, to which carbon taxes apply. This traceability results in flexibility in calculating the marginal reduction costs per tonne of CO_2 for various sources. Commodities (endowments, intermediate goods and final goods) are of two kinds – emission-generating and other. There are economy-wide effects of emissions restrictions, economy-wide effects of self-imposed carbon taxes, and economy-wide effects of technological change.

The data used to calibrate the model are GTAP-4E from 1995. The 'regional' structure comprises: (1) participating Annex I regions, the European Union and other OECD countries; (2) the rest of the world; and (3) non-participating Annex I regions. There are also economies in transition (EIT) and the United States.[34] After discussing extensions to the GTAP-E model to accommodate CDM policies, Wang and Nijkamp focused on procedures and variables within the economic system and emissions in the ecological system affected by introducing such policies. Special algorithms are used to solve interactively for levels of self-imposed taxes that implement the CDM.

Wang and Nijkamp investigate the effectiveness of several policy regimes. Among them are what the authors term: (1) the CDM standard regime; (2) the cap regime; and (3) the USA participation region. In the first and second of these, only the EU and other Organisation for Economic Co-operation and Development (OECD) countries participate as source countries of CDM activities. (The USA does not have an emissions target and does not participate in CDM activities.) Some general findings of Wang and Nijkamp's simulations are that the CDM standard regime, in which emissions are not capped, is more cost effective for Annex I producers than an intra-regional emissions trading regime but that emissions reductions with a cap exceed those obtained in the standard CDM policy regime. They also find that were the USA to participate, favorable impacts on non-Annex I countries would be higher and the percentage reduction in world emissions would be more than double the amount realized when the USA does not participate.

With respect to impacts on regional economies, real GDP decreases for participating Annex I regions but increases for other regions. (All sectors of participating Annex I regions would be faced by declining output.) Except for petroleum in the United States,

the basic energy sectors are faced with lower production, while other sectors gain in production levels.

Roson (2003)

A second use of the GTAP-E model has been made by Roberto Roson to model the economic impact of climate change. Roson's study may be viewed as an effort to overcome limitations on the economic side of integrated assessment models (IAMs), which include overly limited sectoral disaggregation and inadequate articulation of international trade and capital flows. According to Roson, whereas CGE models can help to provide 'a more accurate, realistic and consistent picture of . . . economic systems, their range of applicability is limited by two elements' (p. 1):

1. short- to medium-term horizon for analysis, whereas climate change is a long-term phenomenon; and
2. the environmental dimension is not really present in the model.

Moreover, changes in environmental quality do not enter computations of welfare. Roson and his modeling team at Fondazione Eni Enrico Mattei adopt a two-step approach to attempt to overcome these problems. First, they generate counterfactual equilibria of the world economy and then they conduct conventional comparative statics analyses by simulating shocks. They use impact modules that draw the implications of climate change in different dimensions in order to compute variations in parameters for these experiments. Roson and his colleagues employ an approach developed by Dixon and Rimmer to obtain baseline SAM matrices and model calibrations for future years 2010, 2030 and 2050.[35] 'The idea is to project the structure of the world economy in the absence of any major exogenous shock, including changes in the climate' (p. 3). They focus on supply-side sources of growth, most of which are 'naturally exogenous' in short-run CGE models. In other cases, normally endogenous variables are made exogenous and parameters may be endogenized.

Roson and his colleagues have developed further the GTAP-E model as GTAP-EX by augmenting the industrial disaggregation, especially in the agricultural sector. They obtained estimates of regional labor and capital stocks by employing the G-Cubed model of McKibben and Wilcoxen (1999), a dynamic CGE model of the world economy.[36] They obtained estimates of land endowments and agricultural land productivity from the IMAGE model version 2.2, an IAM with a focus on land use (IMAGE, 2001). Since short-run elasticities are not valid in the long run, prices of natural resources in simulations are pegged to GDP price deflators.

Roson and his colleagues endeavor to move from local to system-wide effects. Because of differences in spatial and sectoral detail, results of impact modules need to be translated in terms of variations of exogenous parameters in the CGE models. Three categories of parameters are affected by this process: productivity shift parameters, resource endowment parameters and structural demand shift parameters.

The researchers found that the effects on health of sea-level changes are non-linear and highly significant in 2050. Health impacts depend on the geographic position of the region. Resource endowments are affected by sea-level rise, as are patterns of land use. Some countries can gain from a sea-level rise. Industries and countries most negatively affected by shocks are those having labor- and land-intensive production processes.

Roson and his colleagues are mindful of limitations of their approach. Because climate change occurs progressively over time, and natural systems interact dynamically with human systems, a fully dynamic IAM is called for. At a sufficient level of disaggregation, however, such a model would be overwhelmingly complex. A more complete understanding of the implications of climate change requires coming to terms with processes by which natural and human systems adapt to the changing climate. On the economic side, the static CGE model would have to be transformed into a dynamic CGE model. This would entail modeling investment behavior and expectations, capital flows, migration and intertemporal budgets. It is unclear how reliable such a model would be in the medium run and even if sufficient information is available to construct and validate such a model. So Roson and his colleagues have elected to continue working in a comparative statics framework.[37]

20.6 SCGE modeling in smaller areas

Notions of general equilibrium in space extend beyond trade. They can also apply to land use, residential location and spatial interaction. In this section we consider the application of SCGE methodologies to the quantitative study of road pricing in a metropolitan area, and the analysis of demographic and land use patterns in smaller areas of analysis.[38]

Sato and Hino (2005)

In light of the success in reducing congestion and transport-related emissions that London and Singapore, among other locations, have enjoyed by introducing road pricing, other metropolitan areas are considering similar measures. In evaluating the potential effects of road pricing, it is important to consider markets for land and labor as well as goods, and to be able to characterize faithfully potential changes in transportation and land use that are likely to ensue from road pricing. Some have turned to computable urban economic (CUE) models to examine changes in land markets in small zones (for example Anas and Xu, 1999; Ueda et al., 2003), as well as more traditional CGE models to make such an assessment. In an applied policy study of the potential effects of road pricing in the Tokyo metropolitan area, Tetsuji Sato and Seiichi Hino develop and simulate an SCGE that differs considerably from those which take the markets to which notions of equilibrium should apply to be limited to the products of industries. The fundamental observation that drives their approach to model specification is that location and travel activities of businesses seem to depend more on functions of businesses than on industrial products.

The basic assumptions of the model are as follows. The economic entities (agents) in the model are households and business firms. Households act rationally, choosing residential location, supplying labor, commuting, consuming goods and services to maximize utility. (The total number of households in the target area is fixed, however, whether or not road pricing is in effect.) Probabilities of selecting particular zones for residence or shopping are given by using the indirect utility function in logit expressions. Businesses are characterized by five functions – administering head office activities, producing goods, retailing of goods, retailing of services and physical distribution – and are also assumed to behave rationally in choosing inputs of labor, land and transport accessibility to maximize profits. Businesses choose to locate in zones where unit costs of operation are low and their probabilities of location choice are also given by logit expressions. An

equilibrium solution to the model comprises equilibrium in markets for goods, labor and land. Wages are determined so that the labor market for the entire metropolitan area is in balance. Transportation is modeled by the traditional four-step approach – trip generation, trip allocation, modal choice and route choice. Revenues from road pricing are rebated to households.

The parameters of expenditure, production and transportation equations of the model were calibrated to conform with various surveys conducted at the metropolitan and national level and engineering studies. The model was validated (quite favorably) for predictive performance against actual population, employment, production, travel origin and destination, and link volume figures.

Simulations conducted included a business-as-usual case in which congestion pricing is not introduced, a second case in which 500 yen per trip is assessed for four inner wards of Tokyo, and a third case in which 500 yen per trip is assessed for 12 wards. As a consequence of road pricing, Sato and Hino find in their simulation results that traffic volumes on ordinary roads in the charge area decrease as users shift to public transport and traffic on ring roads outside the charge areas increases. Population and employers aggregate in the charge area to avoid pricing in coming to central Tokyo. As a consequence, land rents increase in and near the charge area as toll savings are capitalized. The aggregation of employers leads to a realization of agglomeration economies and an increase in gross regional product for the whole Tokyo metropolitan area. The projected toll revenue for the second scenario is 200 billion yen per year and in the third 250 billion yen per year. Significant accompanying reductions in emissions of nitrogen oxides, particulate matter, and carbon dioxide are anticipated, although not reported.

Sue Wing and Anderson (2007)

While the first CGE models were constructed to analyze matters of trade between nations and regions or states, many economic issues cannot be adequately addressed at the national, regional, or even state level. Economically depressed regions seldom coincide with state borders. Analysis of land-use changes needs finer-scaled resolution. Environmental pollution impacts are not contained by state borders. And in transportation studies, origins and destinations are points and impacts of transport infrastructural investments are not contained by state borders. In light of this situation, Ian Sue Wing and William Anderson have endeavored to specify a comprehensive and rigorous framework for sub-state economic analysis. Working in the context of the United States, they have chosen counties over census metropolitan areas (CMAs) as the units of spatial analysis, noting that counties can be aggregated up to CMAs and Bureau of Economic Analysis regions.

Sue Wing and Anderson develop what they term a 'spatially disaggregate' CGE model. The approach they take is to proceed at two levels, developing a county-level model in conjunction with a state-level CGE model. Endogenous variables at the state level serve as boundary conditions at the county level. The county-level model determines a spatial-equilibrium distribution of economic activities and population at the county level within each state. The authors emphasize a rigorous specification on the supply side with an aggregate and simple specification on the demand side. Their model is not yet (at the time of publication) fully operational, but the authors have solved a proof-of-concept model at the state level and conducted empirical estimation of population distribution

mechanisms at the state and county levels. The temporal orientation of the model is dynamic. Their study elaborates the process of model construction.

The first step in developing the framework is to construct a state-level economic simulation model to project trajectories of output, employment, prices and wages by industry in each of the 50 states over the time horizon of 2000 to 2050 by five-year steps. In this model individual industries are treated as representative firms whose investment behavior accords with a Solow–Swan growth model. Output is generated according to a production function with capital, intermediate goods and labor, which is mobile among the states.

Sue Wing and Anderson then build and calibrate a within-period, static spatial price equilibrium (SPE) model of the US economy. In each state there are 20 industries, each producing a single homogeneous commodity according to a constant-returns-to-scale Cobb–Douglas technology from capital, labor and a composite of intermediate goods. Parameters are inferred from the Bureau of Economic Analysis (BEA)'s gross state product accounts, the US SAM, and the 'make and use' table of the BEA's I–O accounts. Output levels at the state level are inferred from value-added, assuming that the economy has the same structure of inter-industry demand at aggregate and state levels. The supply of capital is fixed within a period. The output price offsets the short-run unit cost, under the assumption of competitive equilibrium. The price of intermediates is modeled according to a Leontief specification. The demand industries face is the sum of their own and other industries' intermediate use and final uses by consumers. A representative agent with Cobb–Douglas preferences stands in for final consumers and tax revenues are rebated as a lump sum. Market clearance implies supply equals demand, and within-period equilibrium is closed by defining the labor market at the state level. Sue Wing and Anderson employ Gallin's (2004) model of labor supply.

The dynamic processes by which the economy is advanced through time consist in equations of motion for capital formation and population growth and migration which change in five-year intervals. Boundary conditions from within-period equilibrium solutions determine temporal evolution of spatial patterns of production. The first challenge the authors face is to determine the geographical distributions of growth and decline in industries' capital stocks. Sue Wing and Anderson adopt a recursive dynamic approach. Capital accumulation is determined by the perpetual inventory method and the authors' preferred characterization of investment is an econometrically calibrated investment accelerator model at the industry level; whereas, the evolution of population in each state is modeled as a function of the net growth rate of state populations and immigration and emigration for economic reasons. State population growth rates are modeled using crude birth and death rates. Migration flows are modeled after Greenwood et al. (1991).

To disaggregate economic activity and population to the county level, Sue Wing and Anderson follow a two-stage process of downscaling. First, relationships governing the geographical distribution of output, growth of population and housing stocks, and land-use changes are estimated empirically. Then the resulting set of behavioral equations is solved as a system to obtain the equilibrium spatial distribution. Fifty county-level spatial allocation problems are nested within the state-level spatial allocation problem. Employment and compensation are first allocated industry-by-industry at the county level. Then output is allocated among counties using a logistic sharing rule. Migration flows are allocated on a similar basis.

To determine spatial patterns of housing and land use, Sue Wing and Anderson model a process by which population and income growth induces increased demand for housing and conversion of agricultural land to residential, commercial and industrial uses. In so doing, they obtain expressions which can be used to simulate county-level distributions of economic activity, population and land use.[39]

The model is calibrated as a set of SAMs for the 50 US states. State SAMs are created in a manner different from the usual approach of using multiplier techniques (for example Pyatt and Round, 1985). Sue Wing and Anderson start by creating a national SAM and then disaggregate it to state-level SAMs, replicating the national structure. Recalling that inter-industry demands are assumed to be Leontief, inter-industry demand does not depend on state location. Substitutability of labor and capital implies that factor input ratios can vary in an industry across state lines. State final demands are assumed to reflect patterns in the national SAM, implying that aggregate final use is disaggregated by states' shares of total income. Row and column sums for each industry do not balance at the state level but do so at the national level. Differences indicate magnitudes of net commodity trade flows. Using interstate distances and commodity flow data, interstate trade matrices can be developed.

With disaggregated data, the spatial equilibrium model can be calibrated by setting all prices to unity and solving for technical coefficients that replicate the benchmark data set.[40] Equations of population movement are estimated econometrically using census data and data on socio-economic characteristics. Effects of spatial autocorrelation are captured through spatial lags.

Sue Wing and Anderson view the signal benefit of their research as lying in the creation of 'a simple, transparent, theoretically and methodologically rigorous simulation model that is suitable for a potentially broad range of applications' (p. 284). Among effects that can be examined are how technical progress in industries affects spatial patterns of growth, what the impacts of future state tax policies might be, and what the potential limitations of land conversion are. The framework is beneficial for conducting environmental analysis because it is possible to apply emissions factors to generate spatial patterns of criterial pollutant emissions. The model may also be extended to produce estimates of transportation activity levels and emissions from mobile sources.

By explicitly representing industries' use of intermediate inputs, the model sheds light on the potential for macro-level climate change policy to affect both regional growth and spatial distribution of secondary air pollution benefits from reduced combustion activity.

20.7 New directions in SCGE modeling

The preceding survey should have shown that the label 'SCGE' does not denote a monolith. SCGE models are being constructed and simulated for many different purposes, and models differ on such critical assumptions as returns to scale, competitiveness of markets, treatment of transportation costs, the incorporation of the spatial dimension, temporal orientation, spatial and temporal resolution, elasticities of substitution, rules of closure, and mobility of factors. Still, as work proceeds on many fronts, it may be suggested that there are aspects of SCGE modeling that warrant attention by scholars working within and across different streams of research. We shall close by discussing five: model specification, model calibration, temporal orientation, hypothesis testing, and handling of expectations in policy analysis.

As devices for conducting thought experiments, SCGE models can show us – but show us only – the logical implications of the particular assumptions on which the models are based. These assumptions include the functional forms employed in specifying utility or welfare and production and the values of parameters with which models are calibrated. As noted above, Isard and Azis (1998) have challenged us to use functional forms that are more flexible and overcome the limiting assumptions of common (or symmetric) and constant elasticities of substitution implied by CES aggregator functions. McKitrick (1998) has found that the choice of functional form 'affects not only industry-specific results, but aggregate results as well, even for small policy shocks' (p. 543). Early investigations with flexible functional forms, such as the trans-log or various series expansions, indicated that they could be ill-behaved (irregular) in some areas of the economic region, leading to an emphasis being placed on choosing forms that satisfied certain curvature conditions – for example concavity or convexity, as appropriate (Diewert and Wales, 1987). But Barnett (2002) has shown that one cannot get to global regularity through curvature conditions alone. (For example, elements of the diagonal of a Slutsky matrix may be positive, which violates a necessary condition for regularity of a demand system, even as curvature conditions may be satisfied.) Econometric modelers are accumulating experience in imposing regularity, which is necessary to sanction inferences from model solutions with economic theory (the whole point of a CGE exercise), while gaining sufficient flexibility to overcome limiting assumptions on elasticities.[41] Still, doing so exacerbates already existing degree-of-freedom difficulties with spatial time series. SCGE modelers must come to terms with this trade-off in practice.

As just alluded to, a common lament on the part of SCGE modelers is the lack of sufficient observations to estimate econometrically parameters of SCGE models so that thought experiments can be conducted within robust frameworks whose specifications have been tested against others. Lacking sufficient observations, some may argue – unpersuasively to the minds of econometric critics – that any form of calibration that exploits what data are available is an acceptable solution. One can hear at this point the echo of the exchange between Kydland and Prescott (1996), Hansen and Heckman (1996), and Sims (1996) on calibration versus estimation in macroeconomics. As Sims admonished us then:

> Since there is . . . a tendency in the profession to turn away from all technically demanding forms of theorizing and data analysis, it does not make sense for those of us who persist in such theorizing and data analysis to focus a lot of negative energy on each other. All . . . lines of work . . . are potentially useful, and the lines of work show some tendency to converge. We would be better off if we spent more time in reading each others' work and less in thinking up grand excuses for ignoring it. (Sims, 1996, p. 119)

Until longer spatial time series can be produced, modelers will have to work from both directions – from theory to impose structure, and from econometric methodology to make efficient use of data. Recent work by Clifford Wymer (2006) on the estimation of spatial time-series models provides a positive example, exploring how limited data may be pooled and conditions imposed on variance–covariance matrices, whose validity may be tested post-estimation.

All CGE (including SCGE) models have been widely conceived of as frameworks for conducting numerical comparative-statics exercises, which enable us to gauge how an

economy in equilibrium will be affected by an exogenous shock once it has re-equilibrated. And most of the SCGE models here surveyed are static in temporal orientation. But what we often want to know about – for planning and other purposes – are transition paths between equilibria. While intertemporal analysis has been common in the CGE studies in general, as Pereira and Shoven (1988) and Dixon and Rimmer (2002) attest, it has been less so in SCGE work. Roson (2003) has shown much creativity in giving dynamics to exercises conducted with what are essentially static frameworks to investigate propositions about potential effects of climate change and responding to these effects.[42] But these exercises involve assumptions of plausible steady-state solution paths well into the future for many key driver variables. Modelers have known for many years how to capture endogenous regime switching, for example by employing Dirac functions triggered by endogenous conditions or by taking the approach of Löfgren and Shoven discussed above. But these approaches are rather mechanistic. Our greatest need is for a behavioral theory of spatial dynamic adjustment, especially where path-dependency of development in a spatial economy is at issue. Some answers may lie in adapting PID (proportional, integral and derivative) adjustment approaches, introduced to economics by A.W. Phillips (1954), or in robust control, but these are in many respects just one step less mechanical. Paul Samuelson (1947) has famously observed that 'we damn another man's theory by terming it static, and advertise our own by calling it dynamic' (p. 311). And, admittedly, different models serve different ends. But many of the urgent questions we raise cannot be dismissed by saying dynamic modeling is difficult.

As noted above, Fujita et al. (1999) in closing their book, *The Spatial Economy*, expressed aspirations that work with SCGE models might contribute to making economic geography more of a predictive discipline. Some of us might settle for theorizing leading to more empirically confrontable causal explanations. Indeed, an increasing number of papers are being published in regional science journals in which models generating testable propositions are not themselves being directly confronted by data. Instead (greatly) reduced-form expressions are allowed to stand in their stead. As an alternative, we might use SCGE models to explore how logically consistent empirical findings obtained with partial-equilibrium models are. We need to learn how to triangulate better on economic reality (such as it is) with the tools and capabilities for reasoning we have.

Finally, as Isard and Azis (1998) have urged, we need to relearn some of the positive lessons of the rational expectations revolution of 30 years ago – agents in the private sector anticipate and respond rationally to policy interventions of governments, so their reaction functions need to be embedded in the deep structure of SCGE models if we are to draw appropriate lessons from simulation exercises conducted with them. While Baldwin et al. (2003) have done much to pursue developments in this direction, we need to go further.[43]

Notes

1. The author wishes to acknowledge helpful comments and suggestions from Iwan Azis, Geoffrey Hewings and an anonymous referee.
2. In this chapter, AGE and CGE will be used interchangeably, as will AGIE (applied general interregional equilibrium), ICGE (interregional computable general equilibrium) and SCGE (spatial computable general equilibrium), depending on the terminological preference of the author(s) of the paper under discussion.
3. Going back to Isard (1951) and Moses (1955), 'interregional' and 'multi-regional' have been terms of art connoting differences in approach taken to modeling trade between regions. I will elide the distinction here.

4. This statement pertains to what is widely available in English. There are additional papers in Dutch, Japanese and Swedish that are not accessible to the author.
5. Seldom do modelers working in one of the streams of literature identified above cite work of modelers working in other streams.
6. These include Shoven and Whalley (1984), Periera and Shoven (1988), Robinson and Roland-Holst (1988), Harrison et al. (1993), Kraybill (1993) and Partridge and Rickman (1998).
7. CES is an acronym for constant-elasticity-of-substitution and LES is an acronym for linear expenditure system.
8. Jones and Whalley also report that increasing returns to scale can be incorporated into the production structure of any industry in any region.
9. Jones and Whalley discuss several problems with national and regional reconciliation of the input–output (I–O) data.
10. The 'Hicksian equivalent variation' is the income change needed to compensate the consumer for the price changes, that is, keep the consumer's utility constant (see Varian, 1978).
11. They note that their approach is not without problems.
12. The 'clearing-house assumption' is similar to Leontief and Strout's idea of pools of supply and demand to which regional production and interregional inputs and sales move.
13. A Powell hybrid solution algorithm is used to solve the model.
14. 'A non-linear complementarity problem consists of a system of simultaneous (linear or non-linear) equations that are written as inequalities and linked to bounded variables in complementary slackness conditions. In a mixed complementarity problem, the equations may be a mixture of inequalities and strict inequalities' (Löfgren and Robinson, 1999, p. 3).
15. Rutherford (1995) has since demonstrated that, with algorithms developed more recently to solve non-linear complementarity problems, any neoclassical non-linear demand system can be used.
16. The fixed trade coefficients represent a major bone of contention between Isard (1951) and Moses (1955).
17. The non-linear model they develop can be solved within the GAMS software package alluded to on page 389 of this volume by the PATH and MILES solvers.
18. Friesz, Suo and Westin observe that of the available freight network models, no current model (circa 1998) achieves the simultaneous solution of a general equilibrium model and a freight network model.
19. See note 23 below.
20. There is a trade-off in generality and detail in using supply functions instead of an I–O matrix and in the presence of imperfect competition it is unclear what a supply function means.
21. See note 13 above.
22. Given a subset K of R^n and a mapping $F : K \to R^n$, the finite-dimensional variational inequality problem associated with K is finding $x \in K$ so that $\langle F(x), y - x\rangle \geq 0$ for all $y \in K$, where $<\cdot,\cdot>$ is the standard inner product on R^n. In general, the variational inequality problem can be formulated on any finite- or infinite-dimensional Banach space. Given a Banach space E, a subset K of E, and a mapping $F : K \to E^*$, the variational inequality problem is the same as above where $<\cdot,\cdot> : E^* \times E \to R$ is the duality pairing. (See any text on variational inequalities, for example Nagurney, 1993.) Even if the activity matrix (of inter-industry sales) is not constant, successive linearizations can be performed to create a sequence of VIPs.
23. The notion of a Wardropian user equilibrium is predicated on the assumption that each traveler takes the least-cost route available for travel. All routes connecting an origin–destination pair and carrying some traffic from that origin to that destination have the same cost and all routes which carry no traffic from that origin to that destination do not have a lower cost. See Wardrop (1952).
24. In fact, in spite of Venables and Gasiorek (1996, 1999) having conducted extensive studies with SCGEs, discussed below, FKV chose not to discuss this work in their book.
25. Bröcker (1995) observes that: 'All theoretical ideas used in this paper can be found somewhere in Krugman's recent work' (p. 138).
26. The framework introduced in Bröcker (1995) has been applied to the study of economic integration in Bröcker (1998b, 2000, 2002, 2004).
27. Fan et al. (2000) comment: 'In contrast to the models presented in Fujita et al. (1999), which separately include land rental markets, intermediate inputs, and urban hierarchies, this paper brings together these separate lines of inquiry into a single, unified model' (p. 694).
28. Treyz and Treyz (2002) discuss how aspects of the NEG have been incorporated into the REMI (Regional Economic Models, Inc.) Economic Geography Forecasting and Policy Analysis Model. The changes are manifested in wages, prices, costs and market shares.
29. Cohesion funds are regional funds specifically earmarked for transport and environmental projects in the poorest states of the European Union (http://ec.europa.eu/regional_policy/themes/environ_en.htm).
30. The first two effects are neoclassical quantity effects, whereas the second two are positive feedbacks indicative of cumulative causation.

31. NUTS is an acronym used by Eurostat for 'nomenclature of territorial units'. There are 78 level I territorial units, 201 level II, and 1093 level III. NUTS II is the main analytical level used for European Union regional policy.
32. GTAP is an acronym for the Global Trade Analysis Project. On the general structure of GTAP, see Hertel and Tsigas (1997).
33. Wang and Nijkamp (2007) extends work discussed in Kremers et al. (2002) and Nijkamp et al. (2005).
34. Annex I countries are those listed in the first Annex of the Kyoto Protocol.
35. See also similar work by Giesecke and Madden (2007).
36. G-Cubed is coupled with GTAP-EX because the latter is easier to employ in terms of scale of disaggregation and changes in model equations.
37. Roson notes that he and his colleagues have developed and tested a recursive dynamic version of GTAP-E – called GTAP-ER – and reports that this model produces divergent growth paths for different economies.
38. For another small-area CGE study, which is less spatially explicit, see Learmonth et al. (2007).
39. The key relationships can be solved as a non-linear programming problem in GAMS.
40. The computational model is formulated and solved by using the MPSGE subsystem for GAMS numerical simulation language. MPSGE is an acronym for Mathematical Programming System for General Equilibrium Analysis.
41. Donaghy and Richard (2006) have succeeded in specifying and estimating a globally regular continuous-time system of demand for world money with Divisia data.
42. Following Dixon and Rimmer (2002), Giesecke and Madden (2007) in a different context have also demonstrated how to make regional adjustments to globalization.
43. See also Donaghy and Scheffran (2006) for an explicit dynamic game characterization of a dynamic commodity flow model.

References

Anas, A. and R. Xu (1999), 'Congestion, land use, and job dispersion: a general equilibrium model', *Journal of Urban Economics*, **45**, 451–73.

Ando, A. and B. Meng (2005), 'Transport sector and regional price differentials: a spatial CGE model for Chinese provinces', manuscript.

Armington, P.S. (1969), 'A theory of demand for products distinguished by place of production', *International Monetary Fund Staff Papers*, **16**, 159–76.

Baldwin, R., R. Forslid, P. Martin, G. Ottaviano and F. Robert-Nicoud (2003), *Economic Geography and Public Policy*, Princeton, NJ: Princeton University Press.

Barnett, W.A. (2002), 'Tastes and technology: curvature is not sufficient for regularity', *Journal of Econometrics*, **108**, 199–202.

Batten, D.F. and D. Boyce (1987), 'Spatial interaction, transportation and interregional commodity flow models', in P. Nijkamp (ed.), *Handbook of Regional and Urban Economics*, New York: North-Holland, pp. 357–406.

Bröcker, J. (1995), 'Chamberlinian spatial computable general equilibrium modelling: a theoretical perspective', *Economic Systems Research*, **7** (2),137–49.

Bröcker, J. (1998a), 'Operational spatial computable general equilibrium modeling', *Annals of Regional Science*, **32**, 367–87.

Bröcker, J. (1998b), 'How would an EU-membership of the Visegrád countries affect Europe's economic geography?', *Annals of Regional Science*, **32**, 91–114.

Bröcker, J. (2000), 'Erratum: how would an EU-membership of the Visegrád countries affect Europe's economic geography?', *Annals of Regional Science*, **34**, 469–71.

Bröcker, J. (2002), 'Spatial effects of European transport policy: a CGE approach', in G.J.D. Hewings, M. Sonis and D. Boyce (eds), *Trade, Networks, and Hierarchies: Modeling Regional and Interregional Economies*, Berlin: Springer, pp. 11–28.

Bröcker, J. (2004), 'Regional welfare effects of the European Monetary Union', *Studies in Spatial Development 6*, Hannover: DATAR and Akademie für Raumforschung and Landesplanung, pp. 27–43.

Buckley, P.H. (1992), 'A transportation-oriented interregional computable general equilibrium model of the United States', *Annals of Regional Science*, **26**, 331–48.

Burniaux, J.-M. and T.P. Truong (2002), 'GTAP-E: an energy-environmental version of the GTAP model', GTAP Technical Paper, No. 16.

Diewert, W.E. and T.J. Wales (1987), 'Flexible functional forms and global curvature conditions', *Econometrica*, **55**, 43–68.

Dixit, A.K. and J.E. Stiglitz (1977), 'Monopolistic competition and optimum product diversity', *American Economic Review*, **67**, 297–308.

Dixon, P. and M. Rimmer (2002), *Dynamic General Equilibrium Modeling for Forecasting and Policy*, Amsterdam: North-Holland.
Dixon, P.B., B.R. Parmenter, J. Sutton and D.P. Vincent (1982), *ORANI: A Multisectoral Model of the Australian Economy*, Amsterdam: North-Holland.
Donaghy, K.P. and D.M. Richard (2006), 'Estimating a regular continuous-time system of demand for world monies with divisia data', in M.T. Belongia and J.M. Binner (eds), *Money, Measurement and Computation*, Houndmills: Palgrave Macmillan, pp. 76–103.
Donaghy, K.P. and J. Scheffran (2006), 'A dynamic game analysis of network externality management', paper presented at the 12th International Symposium on Dynamic Games and Applications, Sophia Antipolis, France, July.
Elbers, C. (1996), 'Linking CGE models: modelling the transport sector and spatially homogeneous goods', in J.C.J.M. van den Bergh, P. Nijkamp and P. Rietveld (eds), *Recent Advances in Spatial Equilibrium Modelling: Methodology and Applications*, New York: Springer, pp. 245–60.
Enke, S. (1951), 'Equilibrium among spatially separated markets: solution by electric analogue', *Econometrica*, **19**, 40–47.
Ethier, W.J. (1982), 'National and international returns to scale in the modern theory of international trade', *American Economic Review*, **72**, 389–405.
Fan, W., F. Treyz and G. Treyz (2000), 'An evolutionary new economic geography model', *Journal of Regional Science*, **40**, 671–95.
Friedl, B., C. Schmid and K.W. Steininger (2006), 'Circular causality in spatial environmental quality and commuting: a spatial CGE analysis within a NUTS III Region', mimeo.
Friesz, T.L., Z-G. Suo and L. Westin (1998), 'Integration of freight network and computable general equilibrium models', in L. Lundqvist, L.-G. Mattsson and T.J. Kim (eds), *Network Infrastructure and the Urban Environment*, Berlin: Springer, pp. 212–23.
Fujita, M. and T. Mori (1997), 'Structural stability and urban systems', *Regional Science and Urban Economics*, **27**, 399–442.
Fujita, M., P. Krugman and A.J. Venables (1999), *The Spatial Economy: Cities, Regions, and International Trade*, Cambridge, MA: MIT Press.
Gallin, J.H. (2004), 'Net migration and state labor market dynamics', *Journal of Labor Economics*, **22**, 1–23.
Giesecke, J.A. and J.R. Madden (2007), 'Regional adjustment to globalization: a CGE analytical framework', in R.J. Cooper, K.P. Donaghy and G.J.D Hewings (eds), *Globalization and Regional Economic Modeling*, Heidelberg: Springer, pp. 229–61.
Ginsburgh, V. and M. Keyser (1997), *The Structure of Applied General Equilibrium Models*, Cambridge, MA: MIT Press.
Greenwood, M.J., G.L. Hunt, D.S. Rickman and G.I. Treyz (1991), 'Migration, regional equilibrium, and the estimation of compensating differentials', *American Economic Review*, **81**, 1382–90.
Grossman, G. and E. Helpman (1991), *Innovation and Growth in the World Economy*, Cambridge, MA: MIT Press.
Hansen, L.P. and J.J. Heckman (1996), 'The empirical foundations of calibration', *Journal of Economic Perspectives*, **10**, 87–104.
Harberger, A.C. (1962), 'The incidence of the corporate income tax', *Journal of Political Economy*, **70**, 215–40.
Harker, P.T. (1987), *Predicting Intercity Freight Flows*, Utrecht: VNU Science Press.
Harker, P. and T.L. Friesz (1986a), 'Prediction of intercity freight flows. I Theory', *Transportation Research*, **20B**, 139–53.
Harker, P. and T.L. Friesz (1986b), 'Prediction of intercity freight flows. II Mathematical Fomulations', *Transportation Research*, **20B**, 155–74.
Harrison, G.W., R. Jones, L.J. Kimbell and R. Wigle (1993), 'How robust is applied general equilibrium analysis', *Journal of Policy Modeling*, **15** (1), 99–115.
Hertel, T.W and M.E. Tsigas (1997), 'Structure of GTAP', in T.W. Hertel (ed.), *Global Trade Analysis: Modeling and Applications*, Cambridge: Cambridge University Press, pp. 9–71.
Hirschman, A. (1958), *The Strategy of Economic Development*, New Haven, CT: Yale University Press.
Hitchcock, F.L. (1941), 'Distribution of a product from several sources to numerous locations', *Journal of Mathematical Physics*, **20**, 224–30.
IMAGE (2001), *The IMAGE 2.2 Implementation of the SRES Scenarios*, RIVM CD-ROM Publication 48150818, Bilthoven, the Netherlands.
Isard, W. (1951), 'Interregional and regional input–output analysis: a model of a space economy', *Review of Economics and Statistics*, **33**, 318–28.
Isard, W. and I.J. Azis (1998), 'Applied general interregional equilibrium', in W. Isard, I.J. Azis, M.P. Drennan, R.E. Miller, S. Saltzman and E. Thorbecke (eds), *Methods of Interregional and Regional Analysis*, Aldershot: Ashgate, pp. 333–400.

Isard, W. in association with D.F. Bramhall, G.A.P. Carrothers, J.H. Cumberland, L.N. Moses, D.O. Price and E.W. Schooler (1960), *Methods of Regional Analysis: An Introduction to Regional Science*, Cambridge, MA: MIT Press.
Johansen, L. (1960), A *Multi-Sector Study of Economic Growth*, Amsterdam: North-Holland.
Jones, R. and J. Whalley (1989), 'A Canadian regional general equilibrium model and some applications', *Journal of Urban Economics*, **25**, 368–404.
Kilkenny, M. (1998), 'Transport costs and rural development', *Journal of Regional Science*, **38**, 293–312.
Kim, E., G.J.D. Hewings and C. Hong (2004), 'An application of a transport network multiregional CGE model: a framework for the economic analysis of highway projects', *Economic Systems Research*, **16**, 235–58.
Knaap, T. and J. Oosterhaven (2002), 'The welfare effects of new infrastructure: an economic geography approach to evaluating new Dutch railway links', mimeo.
Koopmans, T.C. (1949), 'Optimum utilization of the transportation system', *Econometrica*, **17** (Supplement), 136–46.
Kraybill, D.S. (1993), 'CGE analysis at the regional level', in D.M. Otto and T.G. Johnson (eds), *Microcomputer-Based Input–Output Modeling: Applications to Economic Development*, Boulder, Co: Westview Press, pp. 198–215.
Kremers, H., S. Wang and P. Nijkamp (2002), 'Modelling issues on climate change policies: a discussion of the GTAP-E model', *Journal of Environmental Systems*, **28**, 217–41.
Krugman, P. (1991), 'Increasing returns and economic geography', *Journal of Political Economy*, **99**, 483–99.
Kydland, F.E. and E.C. Prescott (1996), 'The computational experiment: an econometric tool', *Journal of Economic Perspectives*, **10**, 69–85.
Learmonth, D., P.G. McGregor, J.K. Swales, K.R. Turner and Y.P. Yin (2007), 'The importance of the regional/local dimension of sustainable development: an illustrative computable general equilibrium analysis of the Jersey economy', *Economic Modelling*, **24**, 15–41.
Li, P.-C. and A. Rose (1995), 'Global warming policy and the Pennsylvania economy: a computable general equilibrium analysis', *Economic Systems Research*, **7**, 151–71.
Löfgren, H. and S. Robinson (1999), 'Spatial networks in multiregion computable general equilibrium models', TMD Discussion Paper No. 35, Trade and Macroeconomics Division, International Food Policy Research Institute, Washington, DC.
McKibben, W.J. and P.J. Wilcoxen (1999), 'The theoretical and empirical structure of the G-cubed model', *Economic Modelling*, **16**, 123–48.
McKitrick, R.R. (1998), 'The econometric critique of computable general equilibrium modeling: the role of functional forms', *Economic Modelling*, **15**, 543–73.
Moses, L.N. (1955), 'The stability of interregional trading patterns and input–output analysis', *American EconomicReview*, **45**, 803–22.
Nagurney, A. (1993), *Network Economics: A Variational Inequality Approach*, Dordrecht: Kluwer.
Nerlove, M.D. and E. Sadka (1991), 'Von Thünen's model of the dual economy', *Journal of Economics*, **54**, 97–123.
Nijkamp, P., S. Wang and H. Kremers (2005), 'Modeling the impacts of international climate change policies in a CGE context: the use of the GTAP-E model', *Economic Modelling*, **22**, 955–74.
Nordhaus, W.D. and Z. Yang (1996), 'A regional dynamic general-equilibrium model of alternative climate change strategies', *American Economic Review*, **86**, 741–65.
Nordman, N. (1998), 'Increasing returns to scale and benefits to traffic: a spatial general equilibrium analysis in the case of two primary inputs', CERUM Working Paper No. 7:1998, Umeå.
Partridge, M.D. and D.S. Rickman (1998), 'Regional computable general equilibrium modeling: a survey and critical appraisal', *International Regional Science Review*, **21**, 205–48.
Pereira, A.M. and J.B. Shoven (1988), 'Survey of dynamic computational general equilibrium models for tax policy evaluation', *Journal of Policy Modeling*, **10** (3), 401–36.
Phillips, A.W. (1954), 'Stabilization policy in a closed economy', *Economic Journal*, **64**, 290–323.
Pyatt, G. and J.I. Round (1985), *Social Accounting Matrices: A Basis for Planning*, Washington, DC: World Bank.
Robinson, S. and D.W. Roland-Holst (1988), 'Macroeconomic structure and computable general equilibrium models', *Journal of Policy Modeling*, **10**, 353–75.
Roson, R. (2003), 'Modeling the economic impact of climate change', Foundazione Eni Enrico Mattei, EEE Working Paper Series, N.9.
Round, J.I. (1988), 'Incorporating the interregional, regional, and spatial dimension into a SAM: some methods and applications', in F. Harrigan and P.G. McGregor (eds), *Recent Advances in Regional Economic Modelling*, London: Pion, pp. 24–45.
Roy, J.R. (1995), 'Dispersed spatial input demand functions', *Annals of Regional Science*, **29**, 375–88.
Rutherford, T. (1995), 'Extensions of GAMS for complementarity problems arising in applied economic analysis', *Journal of Economic Dynamics and Control*, **19**, 1299–324.

Samuelson, P. (1947), *Foundations of Economic Analysis*, Cambridge, MA: Harvard University Press.
Samuelson, P. (1952), 'Spatial price equilibrium and linear programming,' *American Economic Review*, **42**, 283–303.
Samuelson, P. (1954), 'The transfer problem and transport cost, II: analysis of effects of trade impediments', *Economic Journal*, **64**, 264–89.
Sato, T. and S. Hino (2005), 'A spatial CGE analysis of road pricing in the Tokyo metropolitan area', *Journal of the Eastern Asian Society for Transportation Studies*, **6**, 608–23.
Scarf, H. (1967), 'The approximation of fixed points of a continuous mapping,' *SIAM Journal of Applied Mathematics*, **15**, 1328–43.
Scarf, H. and T. Hansen (1973), *The Computation of Economic Equilibria*, New Haven, CT: Yale University Press.
Shoven, J.B. and J. Whalley (1984), 'Applied general-equilibrium models of taxation and international trade: an introduction and survey', *Journal of Economic Literature*, **22**: 1007–51.
Shoven, J.B. and J. Whalley (1992), *Applying General Equilibrium*, Cambridge: Cambridge University Press.
Sims, C.A. (1996), 'Macroeconomics and methodology', *Journal of Economic Perspectives*, **10**: 105–20.
Stern, R.M., J. Francis and B. Schumacher (1976), *Price Elasticities in International Trade: An Annotated Bibliography*, London: Macmillan for the Trade Policy Research Centre.
Takayama, T. and G.G. Judge (1964), 'Equilibrium among spatially separated markets: a reformulation', *Econometrica*, **32**: 510–24.
Takayama, T. and G.G. Judge (1971), *Spatial and Temporal Price and Allocation Models*, Amsterdam: North-Holland.
Treyz, F. and G. Treyz (2002), *The REMI Economic Geography and Policy Analysis Model*, Amherst, MA: Regional Economic Models, Inc.
Ueda, T., S. Muto, K. Yamaguchi and K. Yamasaki (2003), 'Empirical evaluation of transport environment policy with a computing urban economic model', paper delivered at the 17th Annual Meeting of the Applied Regional Science Conference, Saitama, Japan (in Japanese).
Varian, H. (1978), *Microeconomic Analysis*, New York: Norton.
Venables, A.J. (1996), 'Equilibrium locations of vertically linked industries', *International Economic Review*, **37**: 341–59.
Venables, A.J. and M. Gasiorek (1996), 'Evaluating regional infrastructure: a computable equilibrium approach', mimeo, London School of Economics.
Venables, A.J. and M. Gasiorek (1999), *The Socio-Economic Impact of Projects Financed by the Cohesion Fund: A Modelling Approach*, Luxembourg: European Commission.
Wang, S. and P. Nijkamp (2007), 'Impact assessment of clean development mechanisms in a general spatial equilibrium context', in R.J. Cooper, K.P. Donaghy and G.J.D Hewings (eds), *Globalization and Regional Economic Modeling*, Heidelberg: Springer, pp. 289–326.
Wardrop, J.G. (1952), 'Some theoretical aspects of road traffic research', *Proceedings of the Institute of Civil Engineering, Part II*, **1**: 325–78.
Wigle, R.M. (1992), 'Transportation costs in regional models of foreign trade: an application to Canada–US trade', *Journal of Regional Science*, **32**: 185–207.
Wing, I.S. and W.P. Anderson (2007), 'Modeling small area economic change in conjunction with a multi-regional CGE model', in R.J. Cooper, K.P. Donaghy and G.J.D Hewings (eds), *Globalization and Regional Economic Modeling*, Heidelberg: Springer, pp. 263–87.
Wymer, C.R. (2006), 'Estimating spatial time-series models', working paper.

21 Modern regional input–output and impact analyses

Jan Oosterhaven and Karen R. Polenske

21.1 Introduction

Economic impact analysis has a long tradition in the input–output (IO) field. A search on Google in May 2007 for impact analyses and IO models produced 1 090 000 records. Many were undoubtedly duplicates or referred to only one or the other of these terms, but we were nevertheless impressed with the proliferation of this technique. Of the various applications of IO models, impact analysis is undoubtedly the most widely used. Many of the early applications estimated economic impacts, but soon analysts were also studying environmental, energy, transportation, land-use and other types of impacts, and these have proliferated greatly beginning in the 1990s. With the recognition of the important worldwide climate change effects, we anticipate that analysts will conduct even more environmental- and energy-impact studies than before.

Underlying the regional analyses is the important basic theory of input–output and socio-economic accounting. We begin by reviewing this basic theory in terms of some of the significant methodological debates that occur. Although not all developments are region-specific, we cover them because regional analysts are beginning to adopt these theoretical advancements in their work. For the applications, we restrict our review to regional and multi-regional impact analyses and the development of computer programming packages that help analysts to conduct such studies quickly.

21.2 Theory of demand-driven IO and SAM impact analysis

One of the most frequently heard criticisms of IO analysis concerns the assumption that the input–output coefficients are constant. By making this assumption, Leontief was able to use data over ten-year and longer periods. Early tests by Carter (1970) and Vaccara (1969) using US national input–output tables for 1939–60 showed that this was not an unrealistic assumption, and Carter (1970) and Strout (1967) found that good output estimates could be achieved by making changes in the input coefficients for only selected sectors, such as chemicals and energy, respectively,

Those of us working at the regional scale, due to lack of data, often had to assume that the national and regional technologies were identical. This seemed less realistic, due to regional product-mix and price differences, than to assume that the actual technology remained constant over time. Rather than to develop new theoretical models to take account of economic geography and spatial accessibility differences, early US regional analysts initially used surveys to obtain the data required for the regional IO tables. By the 1980s, as funding became limited, they devised non-survey methods to estimate regional IO tables, based upon either national technologies or upon regional technologies from a different region. Only since 1990 have new spatial economic theories begun to surface for use with the regional economic impact models. As we show later, the current

regional socio-economic impact analysts are building models to account for changes in both technological and interregional trade relations, thus bringing economic geography theories into prominence.

In practice, non-survey, symmetric IO table (IOT) construction is closely connected with IO model building. In theory, however, they relate to two different operations. Most data construction assumptions use uniform distributions to fill in lacking data in absence of real data. Even when analysts use non-linear gravity- or entropy-maximizing assumptions to construct interregional IOTs (Oosterhaven, 1981a, Appendix; Batten, 1983), their aim is to come as close to the lacking data as possible. The same holds when they use 'industry technology' or 'commodity technology' assumptions to construct symmetric industry-by-industry or commodity-by-commodity IOTs from supply and use tables (Kop Jansen and ten Raa, 1990; ten Raa and Rueda-Cantuche, 2003).

These data construction assumptions differ from the theoretically based behavioral assumptions of fixed intermediate input ratios a_{ij} and fixed primary input ratios c_{kj} used in IO modeling. Notwithstanding the information on changes in these ratios over time, analysts still commonly assume them to be constant when modeling the future or when estimating impacts of specific exogenous changes with the Leontief model. One main justification is that, for closed economies, the analyst can theoretically derive fixed ratios by assuming profit-maximizing firms that produce total output x_j with a Walras–Leontief production function under constant returns to scale:

$$x_j = \max\,(z_{ij}/a_{ij}, \text{ with } i = 1, ..., I, v_{kj}/c_{kj}, \text{ with } k = 1, ..., K), \tag{21.1}$$

and that firms sell this output on markets with full competition. This combination of assumptions assures that, irrespective of the relative prices of intermediate and primary inputs, cost minimization with a given total output x_j always leads to fixed input ratios and to the well-known IO demand functions for intermediate and primary inputs:

$$z_{ij} = a_{ij}\, x_j \text{ and } v_{kj} = c_{kj}\, x_j \quad \text{or in matrix notation: } \mathbf{Z} = \mathbf{A}\,\hat{\mathbf{x}} \text{ and } \mathbf{V} = \mathbf{C}\,\hat{\mathbf{x}} \tag{21.2}$$

where $\hat{\mathbf{x}}$ denotes a diagonal matrix, and where the combined column sum of **A** and **C** equals one, that is, **i' A + i' C = i'**. Note that equation (21.2) implies full economic complementarity of all inputs.

Adding the definition of total demand and assuming that the supply of output always follows total demand gives:

$$x_i = \Sigma_j\, z_{ij} + \Sigma_q f_{iq} \text{ or in matrix notation: } \mathbf{x} = \mathbf{Z}\,\mathbf{i} + \mathbf{F}\,\mathbf{i} \tag{21.3}$$

Solving equations (21.2)–(21.3) for exogenous final demand **f = F i**, leads to the well-known solution of the IO model for any aggregate impact variable *v*, such as total regional or national value-added, employment, energy use or CO_2 emissions:

$$v = \mathbf{c'}\,\mathbf{x} = \mathbf{c'}\,(\mathbf{1}-\mathbf{A})^{-1}\mathbf{f} \tag{21.4}$$

in which **c'** indicates a row with value-added, employment, energy use, or CO_2 emission coefficients per unit of output. Note that equation (21.3) also implies that any exogenous

or endogenous change in demand is fully met by the appropriate supply of intermediate and primary inputs. Analysts who use the standard IO model thus assume that there are no supply restrictions. Therefore this IO model is best labeled as a demand-driven quantity model.

This foundation of equation (21.4) in production theory becomes insufficient if we move from a closed economy to an open economy. Consider the closed world economy as a system of R open regional economies, then (21.4) mathematically also describes the interregional IO model (Isard, 1951) as well as the multi-regional IO model (Chenery, 1953; Moses, 1955; Polenske, 1980), but then the vectors **c**, **x** and **f** all have dimension IR and the matrices **I** and **A** have dimension IR x IR. The most important difference with the closed economy model is that the intermediate input coefficients now also get two dimensions:

$$\text{interregional IO model: } a_{ij}^{rs} = t_{ij}^{rs} a_{ij}^{\cdot s} \qquad \text{multi-regional IO model: } a_{ij}^{rs} = t_{i\cdot}^{rs}\, a_{ij}^{\cdot s} \tag{21.5}$$

with the $\cdot$ indicating the summation over the relevant index. Equation (21.5) explicitly shows that the interregional and the multi-regional input coefficients a_{ij}^{rs} represent the product of real technical coefficients $a_{ij}^{\cdot s}$ from production function (21.1), and cell-specific *trade* coefficients t_{ij}^{rs} in the case of the interregional input–output (IRIO) model, or (row) aggregate trade coefficients $t_{i\cdot}^{rs}$ in the case of the MRIO model. In the case of the single-region IO model, the intra-regional trade coefficients t_{ij}^{rr} and $t_{i\cdot}^{rr}$ are better known as, respectively, cell-specific or aggregate, self-sufficiency ratios or regional purchase coefficients (RPCs; Stevens and Treyz, 1986).

The theoretical foundation for assuming the trade coefficients to be fixed is less convincing than that for the technical coefficients. The analyst may assume that the output of, for example, agriculture is a different product in each different region. The trade coefficients will then get a technical character and will be fixed for the same reasons as the technical coefficients. As each cell then relates to different goods, this assumption fits best with the IRIO model. The analyst may also assume that the products of, again for example, agriculture in different regions are close substitutes for each other. The trade coefficients will then only be fixed for as long as the relative prices of agricultural output from different regions remain unchanged. As relative prices will influence all trade coefficients along a row of the IO table in the same manner, this assumption fits best with the MRIO model.

In interregional impact studies, equation (21.4) is disaggregated by regions to allow for the separate estimation of intra-regional impacts and interregional spillover effects (Miller and Blair, 1985). When intra-regional impacts from the single-region model are compared with those from the interregional model, the latter will be larger. This difference is caused by interregional feedback effects. These link the imports of the home region to the output levels of other regions, which are linked back to the intermediate exports of the home region. As a consequence, exogenous final demand will be smaller than in the single-region model, as the export of intermediates becomes endogenous. Of course, when the smaller exogenous final demand is multiplied by the larger multipliers that include the interregional feedbacks, the resulting endogenous employment and value-added will be the same. Interregional feedbacks are found to be relatively large (5–15 percent) for regions within well-integrated large conurbations and smaller ($<$5 percent) for relatively isolated regions (Oosterhaven, 1981a).

The size of the feedbacks and this difference become larger when the standard IO model is extended with a consumption function for labor incomes, as labor incomes earned in the home region will spill over into other regions and feed back into the home region through the additional mechanisms of interregional commuting and interregional shopping (Madsen and Jensen-Butler, 2005). There is a whole family of demo-economic extensions of the basic Type I model into Type II, III, and so on, regional IO models (Batey, 1985). However, the important distinction between increases in labor incomes accruing to resident workers (intensive income growth), new labor incomes accruing to migrants and unemployed (extensive income growth), and the loss of benefits of formally unemployed (redistributive income growth) can only be modeled properly if levels of economic activity are explicitly distinguished from changes therein (Oosterhaven and Folmer, 1985).

With levels and changes in levels distinguished, the interregional Type III variant of equation (21.4) becomes:

$$\Delta v = \mathbf{c'}\,\Delta\mathbf{x} + \Delta\mathbf{c'}\,\mathbf{x}_{-1} = \mathbf{c'}\,(\mathbf{I} - \mathbf{A} - \mathbf{Q_w}\,\mathbf{W}\,\hat{\mathbf{c}}_\mathbf{w} + \mathbf{Q_u}\,\mathbf{U}\,\hat{\mathbf{c}}_\mathbf{u})^{-1}\,\Delta\mathbf{f} + \Delta\mathbf{c'}\,\mathbf{x}_{-1} \qquad (21.6)$$

If Δv represents, for instance, the system-wide change in employment, $\Delta\mathbf{c'}$ represents the *IR*-row with decreases in employment coefficients due to nominal labor productivity increases, and $\mathbf{x}_{-1}$ represents the output impact of the combined lagged endogenous and exogenous variables. Furthermore, $\mathbf{Q_w}$ and $\mathbf{Q_u}$ represent the consumption expenditure on products from sector *i* in region *r*, respectively, per working resident and per unemployed resident in region *s*. The interregional diagonal blocks of $\mathbf{W}$ (with $\sum_r w_j^{rs} = 1$) represent the shares of the new jobs in sector *j* in region *s* directly or indirectly taken up by residents of region *r*, and the comparable blocks of $\mathbf{U}$ (with $\sum_r u_j^{rs} < 1$) represent the shares of the new jobs in sector *j* in region *s* directly or indirectly taken up by residents of region *r* that were formally unemployed. Finally, the diagonal matrices $\hat{\mathbf{c}}_\mathbf{w}$ and $\hat{\mathbf{c}}_\mathbf{u}$ represent the per unit labor incomes and the per unit lost unemployment benefits, respectively. Typically, $\mathbf{W}$ and $\mathbf{U}$ are determined by means of an IO vacancy-chain sub-model. With the unemployment benefits of the Netherlands, Type III impact multipliers $\mathbf{c'}(\mathbf{I} - \mathbf{A} - \mathbf{Q} + \mathbf{Q^u})^{-1}$ from equation (21.6) move between 35 percent and 60 percent of the difference between Type I multipliers $\mathbf{c'}(\mathbf{I} - \mathbf{A})^{-1}$ and Type II multipliers $\mathbf{c'}(\mathbf{I} - \mathbf{A} - \mathbf{Q})^{-1}$ (van Dijk and Oosterhaven, 1986).[1]

The distinction between demo-economic models and social accounting models (SAMs) is not large in the sense that a SAM may be interpreted as a demo-economic model with a more elaborate disaggregation of the household sector. It is, however, more fundamental in that SAMs invariably start with defining the underlying accounting framework explicitly, and then derive their impact multipliers more or less directly from it. Pyatt (2001) uses this approach to show that sectorally and regionally disaggregated Keynesian income multipliers (Miyazawa and Masegi, 1963; Miyazawa, 1976) can be viewed as special cases of various SAM multipliers (Pyatt and Thorbecke, 1976; Pyatt and Round, 1979). The core difference is that Miyazawa multipliers relate the factor (capital versus labor) generation of income by sector and region, directly to the spending of that income on products from different sectors and regions. SAM multipliers are more general in that they add the redistribution of income by different institutions in-between the generation and the spending of income to the IO model. As a consequence, SAMs are better suited to study the impact of policy instruments on the distribution of income and poverty.

The obvious next question is: how far should an analyst endogenize the various components of final demand? Studying the regional impacts of plant close-downs with a SAM, Cole (1989, 1997) advocates the fullest possible closure of the single-region model to capture all possible short- and long-run impacts. Government expenditure is made dependent on tax income, investment expenditure on the operating surplus, and regional exports on regional imports. This led to a major debate with Jackson et al. (1997) about zero exogenous demand and infinite multipliers, which was concluded by Oosterhaven (2000). Analysts cannot endogenize interregional feedbacks consistently without specifying the full interregional model, in which case the full closure of the rest of the model indeed results in zero exogenous demand. As a consequence, such a SAM may no longer be used to evaluate the impacts of changes in demand, because these have become endogenous.

The danger of overestimating impacts also occurs when total value-added or employment of an existing firm or sector is multiplied by that firm's or sector's Type II normalized value-added or employment multiplier, $\mathbf{c'}\,(\mathbf{I} - \mathbf{A} - \mathbf{Q})^{-1}\,\hat{\mathbf{c}}^{-1}$, to indicate its economic importance for the economy at hand. Of course, this is a misuse of impact analysis for public-relations purposes. Formally, an analyst may only multiply an IO multiplier by exogenous final demand and never by endogenous value-added or employment. Imagine that the average normalized employment multiplier equals 2 and that the analyst would apply this way of estimating the impact to all sectors; then, the predicted size of the total economy would be twice its actual size. One solution is to correct the calculated 'gross impact' of a certain sector with the part of that sector that is endogenously dependent on the rest of the economy in order to obtain the net impact of that sector (see Oosterhaven et al., 2003, for energy distribution). A second solution is to define a net multiplier that may be multiplied with a sector's total employment or output, such that the weighted average of all sectors' net multipliers equals one (Oosterhaven and Stelder, 2002).[2]

More generally, the fullest possible closure of standard IO models exacerbates the theoretically already problematic one-sidedness of that model. Possible shortages on local labor markets, price and wage reactions, and the pressure to develop new markets and new products will, for example, reduce the demand-driven quantity impact of a plant close-down. In such cases, endogenizing price effects and supply reactions may be more important than a full closure of the demand side, that is, if analysts are interested in the best estimate of the real impacts instead of maximum multipliers.

21.3 Theory of price and supply-side impact analysis

An analyst may use the less well-known dual of the Leontief model to estimate endogenous output price impacts $\mathbf{p}$ of the exogenous primary input prices $\mathbf{p}_v$ (Leontief, 1951). The dual, however, cannot be used to model the effects of price changes on quantities, as follows from the solution of the standard (Type I) price model (Schumann, 1968):

$$\mathbf{p'} = \mathbf{p}_v\mathbf{'}\,\mathbf{C}\,(\mathbf{I} - \mathbf{A})^{-1} \tag{21.7}$$

In equation (21.7), the K exogenous, capital, labor, and import price changes $\mathbf{p}_v$ are directly passed on via $\mathbf{C}$ into output price changes, and are then indirectly carried forward further via the rows of $\mathbf{A}$ into equilibrium total output price changes. Therefore, equation (21.7) is best characterized as a cost-push price model.

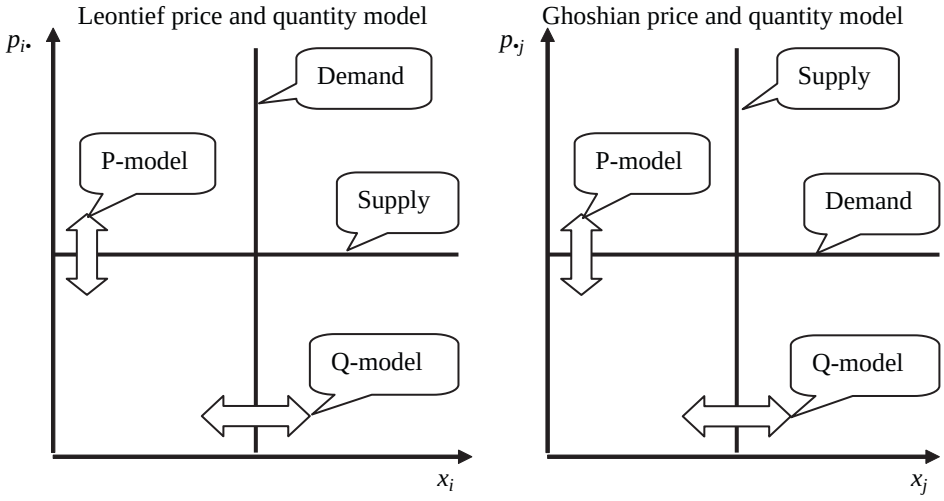


Figure 21.1 The functioning of markets in Leontief and Ghoshian IO models

In the left-hand panel of Figure 21.1, we show the relationship between the demand-driven quantity model and this cost-push price model for an individual IO market. The demand curve (driven by the quantity model) shifts left and right along the perfectly elastic supply curve (no shortages). Independently, the supply curve (driven by the price model) shifts up and down along the perfectly inelastic demand curve. This IO price model, *inter alia*, has been used to estimate the price effects of pollution abatement (Giarratani, 1974) and the effects of energy price rises in a multi-regional input–output (MRIO) model (Polenske, 1979).

When the interregional quantity model is extended with a consumption function for labor incomes, its dual price model of course changes analogously. Wages become endogenously determined by the prices of consumption goods, and the remaining exogenous primary input prices get multiplied with larger Type II cost-push multipliers resulting in the same equilibrium prices for total output (see Oosterhaven, 1981b, for interregional wage and price impacts of the oil price hike).

Also disregarding the price-induced impact of supply on demand, Giarratani (1976), Davis and Salkin (1984) and Chen and Rose (1985) have used the supply-driven IO model, developed by Ghosh (1958), as a direct way to model the impacts of natural resource shortages on output. In every respect, this quantity interpretation of the Ghosh model represents the pure opposite of the demand-driven quantity model, as follows from its solution for endogenous final demand (Oosterhaven, 1996):

$$\mathbf{F} = \hat{\mathbf{v}}\,(\mathbf{I} - \mathbf{B})^{-1}\,\mathbf{D} \tag{21.8}$$

In equation (21.8), exogenous primary supply $\hat{\mathbf{v}}$, without further need of intermediate inputs, directly leads to equally large total inputs $\hat{\mathbf{x}}$ (the direct forward effect), which is distributed to purchasers according to *IxI* fixed intermediate output coefficients, $b_{ij} = z_{ij}/x_i$, and *IxQ* fixed final output coefficients, $d_{iq} = f_{iq}/x_i$. The intermediate outputs, without further need of primary or intermediate inputs, lead to equally large total inputs (the first

round indirect forward effect), which are again distributed to intermediate and final purchasers according to **B** and **D**, and so on.

Originally, Ghosh formulated his model not in terms of quantities, but in terms of values. Dietzenbacher (1997) proved that a value interpretation can be formulated such that it is equivalent with the Leontief cost-push price model. Quantities remain unaffected, and analysts may evaluate final output price impacts of exogenous primary input prices in terms of values instead of in terms of prices. With this interpretation, the row sums of the Ghosh-inverse $(\mathbf{I} - \mathbf{B})^{-1}$ may still be used as a descriptive statistic, measuring the size of a sector's forward linkages, just as the column sums of the Leontief-inverse $(I - A)^{-1}$ are used to measure the size of the backward linkages.

After the criticism by Oosterhaven (1988), the quantity interpretation of the supply-driven IO model is no longer used, as it theoretically allows cars to drive without gasoline and factories to work without labor. Only by intelligently combining processing coefficients (= inverses of the technical coefficients $a_{ij}^{\cdot r}$ and $c_{kj}^{\cdot r}$) with intermediate output or allocation coefficients (b_{ij}^{rr}), while adapting regional purchase coefficients (t_i^{rr}) to accommodate for import and export substitution (Oosterhaven, 1988), may the forward effects idea of the supply-driven IO model still be used (see Cartwright et al., 1982, for a nuclear disaster application, and Oosterhaven, 1981a, for a land-reclamation application).

Interestingly, the dual of the Ghosh quantity model (Davar, 1989) has not been used for price impact studies yet, whereas this is clearly possible, as follows from its solution:

$$\mathbf{p} = (\mathbf{I} - \mathbf{B})^{-1}\,\mathbf{D}\,\mathbf{p_f} \tag{21.9}$$

In equation (21.9), the Q prices for each column with homogenous final inputs ($\mathbf{p_f}$) are exogenous, whereas the I prices for each column with homogenous sectoral inputs ($\mathbf{p}$) are endogenous. Any increase in, for example, the unit price for exports p_q leads to a direct increase in the unit revenues of each sector i with $d_{iq}\,p_q$. Under full competition this increase is entirely passed on into the price of the single homogenous input of sector i. In the next round, this price increase leads to an increase in the unit revenues for all sectors j with $b_{ji}\,p_i$, which is further passed on again via **B**, and so on. Naturally this model may best be characterized as the demand-pull price model.

The right-hand panel of Figure 21.1 shows how the Ghoshian price and quantity model 'interact'. In the quantity model, the supply curve shifts left and right along a perfectly elastic demand curve, not causing any price reaction (consumers consume whatever is supplied at the going price). Independently, in the price model, the demand curve shifts up and down along the perfectly inelastic supply curve, not causing any quantity reaction. According to Oosterhaven (1996) the demand-pull price model is less implausible than its companion supply-driven quantity model. But it is clear from Figure 21.1 that the more plausible Leontief price and quantity model also may be labeled as 'special'.

In fact, both sets of IO models represent extreme cases of the general equilibrium model. Clearly, implementing the latter model at the combined inter-sectoral and inter-regional level is more complicated and far more data-demanding than the comparable IO model. For this reason, most developments in impact analysis seek to modify the basic Leontief model by introducing more flexible (for example translog) production functions for capital, labor, and intermediate input (for example the KLEM production function, that is, Kapital Labor Energy and Materials), and by introducing econometrically

estimated consumption, investment and export functions, while sticking to the Leontief specification for the matrix of intermediate demand only.

21.4 Developments in IO data construction and model applications

Spatially, analysts have constructed both regional and multi-regional IO impact models.

Regional IO table construction and applications

At a regional scale, we have progressed relatively slowly from the initial impact studies, which included Moore's (1955) classic article on 'Regional economic reaction paths', to the well-explained inter-industry model for Utah by Moore and Peterson (1955), in which they used a supply-demand pool procedure to estimate the regional transactions, to Hirsch's (1959) St Louis model for which he collected data from company records and built a detailed export sector. Miernyk's work in the nine Colorado river basins (with Udis), and in the state of West Virginia, certainly contributed greatly to the collection of regional data through surveys in the United States. In his easy-to-read book *The Elements of Input–Output Analysis* Miernyk (1965) gave many useful guidelines for survey design and model building, including how to select 'best-practice' (most technologically advanced) firms. He also started the use of the terms 'Type I' and 'Type II' multipliers. The most extensive and longest-lasting collection of regional data through surveys is by the Washington State researchers in a tradition dating back to 1963 and continuing to today.[3] That state table is used as the standard by many US regional analysts.

Since the 1970s, most IO tables for other US regions have been estimated using various short-cut techniques, such as multiplying national technical coefficients by regional location quotients, primarily because of the high costs involved in conducting surveys. These non-survey techniques, however, produce relatively inaccurate regional tables (see Polenske, 1997, for an evaluation). Moreover, by using location and other quotients, most methods implicitly maximize intra-regional transactions and minimize regional imports in one way or another. Consequently, all regional multipliers from such tables have a systematic upward bias, even when analysts claim that there is relatively little cross-hauling (as West, 1990, p. 108, does for Australia). In more densely populated and diversified economies, however, cross-hauling is the rule, even more so when commuting across regional boundaries is important and Type II multipliers are concerned (Oosterhaven, 1981a).

Many US regional analysts use the Washington State table either to create input–output tables for their own region or to test for differences in the national and state coefficients. Analysts could reduce errors to less than 10 percent only by sectoral aggregation or by using exogenous information on more than 60 percent of the non-zero cells (Polenske, 1997). The latter observation stresses the conclusion of Lahr (1993) that hybrid methods should be preferred to simple non-survey methods.

A hybrid Dutch regional IO table, for example, resulted from a survey into four tourist-specific branches in the province of Drenthe. The survey regional import coefficients of the tourist branches were applied to the national technology coefficients for the other sectors. The complete IO table resulted by combining the survey and non-survey columns into a single table. A comparison with the multipliers from a survey-based bi-regional table showed differences in the indirect part of the tourist branches' multipliers of only 5–10 percent, whereas the differences in the indirect parts of the

other multipliers ran from –50 percent to +65 percent, with some outliers beyond that (Spijker, 1985). We believe that this shows that both Bourque (1990) and Beemiller (1990) are right in their discussion in the *International Regional Science Review*; Bourque in his rejection of RIMS's non-survey alternative for the Washington State IOT and Beemiller in his claim that combining direct information for the sectors of an impact study with a non-survey IOT produces sufficiently accurate estimates for most practical impact questions.

The Washington State IO model is widely used for a diverse set of economic impact analyses. Bill Beyers, for example, used it to estimate the economic impacts of arts and cultural organizations on the Washington economy; the impacts of building and operating the new Seattle symphony hall; the impacts of the Mariners Baseball Team; and those of the Fred Hutchinson Cancer Research Center. As indicated above, the Washington IOT is extremely well known among US regional planners, many of whom have used it to calculate the accuracy of their own tables, and/or to estimate state tables based upon the Washington State table.

For the United States, two options seem relevant. First, regional technologies are sufficiently diverse that regional analysts should be arguing for funds to conduct surveys to derive their tables based upon actual data for that region. Second, analysts should be devising new methods of estimating regional tables from the available data. Neither effort is occurring, with the exception of the MITER accounts being developed by the US Department of Agriculture (USDA) and Massachusetts Institute of Technology (MIT) staff, described later on. We note that the construction of tables through survey methods requires at least $200 000 per state. If each of the 51 US regions were to construct such a table, the total would be over $10 million. Thus, the rationale for assembly of a complete set of state tables by a central government unit is obvious.

Elsewhere, developments have been only partially comparable. Dutch regional IO analysis up to the 1970s may be summarized as running 'from regional tables with only limited information used for primarily descriptive purposes towards ideal interregional tables used for analytical purposes, such as estimates of economic impacts, experiments with programming models and building full forecasting models' (Oosterhaven, 1981a, p. 23). As opposed to the Dutch tables constructed in the 1970s and also different from most tables constructed in the United States and in Australia (West, 1990), the Dutch bi-regional tables of the 1980s are mainly based on surveys of export coefficients instead of import coefficients. This change in trade survey strategy also led to the need of an explicit domestic intermediate and final sales table (Boomsma and Oosterhaven, 1992).

This strategy change was the outcome of earlier experiences that showed that firms, as a rule, are better informed about the destination of their output than about the origin of their inputs. This is especially the case if there are many different inputs and/or if inputs are purchased through wholesale or retail channels. Firms are only well informed about the real origin if they deal with one or a few dominant purchases directly from the producer of the inputs. On the sales side, however, firms may lack the necessary information on the spatial destination of their sales only if they primarily sell through wholesale firms. But even then, they appear to be better informed about their sales than about their purchases through wholesale firms.

We will not deal with regional sector-specific models. Although many exist, the issues are similar to those for regional models.[4]

Multi-regional and interregional IO tables and applications

Since the 1960s when Isard (1960) published his text on regional analysis, many IO analysts have considered that interregional input–output (IRIO) or multi-regional input–output (MRIO) tables are needed to derive and compare the direct and indirect economic impacts across regions. As globalization of the markets increases, the need for such tables by analysts increases. In the 1960s, probably as a result of interchanges with the Japanese, who were gathering extensive IO data, Isard set forth an IRIO set of accounts in which each IO coefficient was specified in terms of the region and sector in which the input originated and the region and sector in which it terminated. The Japanese used that interregional accounting framework to construct the first IRIO set of accounts for nine regions and ten commodities for the 1960 and 1963 Japanese economy.

Leontief felt that such detail was not theoretically justified for the technological inputs, because when an engineer makes a widget, the region in which the input was produced should not matter. In his national–regional model, therefore, he did not specify the region of origin of any input (Polenske, 1999). Both Isard and Leontief were correct, in their own way. On the one hand, for determining the technology (production function) of a sector in a particular region an engineer or economist does not need to know the region from which the inputs come. The analyst only needs to know the production technology being used in each region for each sector. Leontief's national–regional and MRIO accounts (Leontief and Strout, 1963; Polenske, 1970, 1980, 1995) therefore require considerably less information, thus cost less and take less time to construct than Isard's IRIO accounts. On the other hand, from a transportation planner's perspective, the regional origin of the input is critical for transportation planning, as well as for tracing the supply chains of commodities (Polenske and Hewings, 2004).

A major innovation with the latest MRIO accounts (Canning and Polenske, forthcoming) is that the data will not only represent actual data from the census, as the first US MRIO accounts did,[5] but the data are being collected from electronic files at the state and county levels, using algorithms that will make it relatively easy to construct accounts in the future, thus greatly reducing the time required to construct these accounts. Special attention is being paid to estimating the suppressed data from the census. Most importantly, the data will be freely available. This is an important consideration given that the data in most other models are available only by paying vendors in conjunction with buying a model.

We note that several US groups, such as IMPLAN (Impact Analysis for PLANning), REDYN (Regional Dynamics), and NIEMO (National Interstate Economic Model) are constructing or have constructed MRIO accounts, but as far as we can determine, they are not making their data freely accessible. IMPLAN has been compared with REMI (Regional Economic Models, Inc.) and with RIMS-II (Regional Input–output Multipliers, version II) by a number of analysts, with the finding that after adjusting for differences in the models, the multipliers are relatively similar. Richman and Schwer (1993, p. 143) for example found 'that apparent changes in the multipliers in each model result from undocumented or poorly documented changes in the vendor default values of the available options for calculating the multipliers, not from structural changes in the models'.

According to the REDYN vendors, their model is 'more flexible and versatile' than other commercially available models, but so far no independent comparative evaluation

is available. The REDYN model uses the North American Industrial Classification System at the five-digit detail level (703 sectors), identifies the more than 180 commodities consumed and produced by these sectors, and provides forecasts for over 800 occupations. They obtain the underlying data with the use of County Business Patterns from the US Bureau of the Census, and the Regional Economic Information System from the US Bureau of Economic Analysis. The trade flows are provided for five transportation modes at the county level for 3100 counties. Former REMI staff are constructing REDYN, with the intent to provide on-line capabilities with up-to-date information.

NIEMO is a combined supply–demand-driven IO model developed by analysts at the University of Southern California's National Center for Metropolitan Transportation Research and the Center for Risk and Economic Analysis of Terrorist Events, as part of their research at the Homeland Security Center for Excellence. They constructed the 47-sector model for 52 regions (50 states, Washington, DC, and Rest-of-the-World). The theoretical nature of the iteratively solved model is not entirely clear. One of its main uses so far has been to determine the economic impacts on these regions resulting from a hypothetical terrorist attack on the three major US ports of Long Beach, Newark and Houston (Park et al., 2007) They made use of IMPLAN for the technology part of the model and developed interregional shipments for the 47 sectors, using the doubly constrained Fratar model (Richardson et al., 2005).

The development of several competitive multi-regional IO models is a sign that such analytical impact-assessment tools with real data are in great demand.

Japan is the main country with full interregional IO accounts, which are available for 1960 and every five years since, for 42 prefectures, as well as originally for nine regions, for ten sectors (Abe, 1986). They have set the standard for such detailed assembly of data. Currently, China has MRIO accounts for 1987 constructed over a five-year period in collaboration with the Japanese for seven regions and nine sectors (Ichimura and Wang, 2003; Okamoto and Ihara, 2005). Analysts are using these tables to examine many important regional topics in the rapidly growing economy of China, including factors creating the regional differentials in income. Although the types of studies completed so far are typical economic impact assessments, the availability of improved and detailed current regional data in the near future will provide important possibilities for new analyses on energy, environment, transportation, foreign trade and other current topics.

Since 1995, several countries, such as Canada (Siddiqi and Salem, 1995), the Netherlands (Eding et al., 1999) and Finland (Piispala, 2000), have embarked upon the construction of multi-regional supply and use (commodity-by-industry) tables. Note that not all of these rectangular accounting schemes have a straightforward one-to-one relation with the symmetric IRIO and MRIO models discussed in section 21.2 (Oosterhaven, 1984).

21.5 Modern impact analysis

Today, many analysts are further developing social accounting models (SAMs) and intersectoral computable general equilibrium (CGE) models, and they extensively use such models for economic impact analyses. Regional analysts are conducting numerous economic impact analyses at multi-regional levels in the United States, most of which use one of three computer programs available from commercial vendors: REMI, IMPLAN and RIMS-II, with the matrix multipliers furnished by the US Bureau of Economic Analysis.

REMI is an eclectic multi-regional model that combines economic base, input–output, computable general equilibrium, and econometric methods, and was first put forward for Massachusetts by Treyz et al. (1980). ECOTEC REMI is the daughter model for the United Kingdom, and REMI-NEI is the daughter model for the Netherlands. REMI is a sophisticated regional model with numerous policy variables. Analysts use it to determine environmental pollution impacts, regional impacts of transport infrastructure, including airports, and studies of many large investment projects. Latest versions of REMI include job-accessibility measures and spatial-agglomeration effects.

Mourouzi-Sivitanidou and Polenske (1988) conducted one of the earliest evaluations and comparison of these and other impact-assessment models.[6] One of the most extensive and longest uses of REMI is in Los Angeles by the South Coast Air Quality Management District (SCAQMD), which has used it since 1989 for the determination of job impacts of its rules and regulations, and since 1994 for its tradeable permit Regional Clean Air Management program. The SCAQMD uses it to study job impacts in a four-county (Los Angeles, Orange, Riverside, San Bernardino) region in southern California. Polenske et al. (1992) conducted an extensive 13-month evaluation of the use of REMI by the SCAQMD, determining that as of 1992, the SCAQMD had state-of-the-art impact assessments with the use of REMI. The team provided over 30 recommendations for possible improvements both to the model and to the other evaluation methods being used, many of which have been implemented.

In addition, other readily available regional modeling packages for the United States include IMPLAN and RIMS-II. Each of these packages costs money, with RIMS-II being the least expensive and REMI the most expensive. REMI, however, is the only regional economic model that analysts can use for forecasting over 10–20 years. IMPLAN and RIMS-II have more sectors than REMI, but they can only be used for comparative-static impact assessments.

IMPLAN was started in the 1970s in a combined effort by Lofting at the Berkley Lawrence Livermore National Laboratory and Alward at the US Forestry Service. In 1993, the Minnesota IMPLAN Group was founded by Lindall and Olson, based upon work at the University of Minnesota. At the forest service starting in the 1970s, Alward and others developed Micro-IMPLAN, an input–output impact model specifically designed to meet the US Department of Agriculture (USDA) Forest Service needs. Results for four years (1990, 1985, 1982 and 1977) are reported by Shields et al. (1995).[7] The major characteristics of IMPLAN Version I, which is no longer used, were: (1) industry-by-industry accounting; and (2) supply–demand pooling technique for trade estimations. IMPLAN Version II is the basis for all subsequent IMPLAN software, including the current (2006) release of MicroIMPLAN Rel 91–F. Its principal characteristics are: (1) rectangular accounting (that is, make-and-use matrices); and (2) regional purchase coefficients (RPCs) for trade estimation.

IMPLAN differs from REMI in several ways. First and perhaps most importantly, IMPLAN is a static IO impact model, based on the latest national IO table (as of 2008 for the year 2002). In contrast, REMI is a dynamic input–output, econometric forecasting and simulation model, in which Bureau of Labor Statistics forecasts of the input coefficients are used to obtain regional technology forecasts for up to 20 years. The 2008 version of IMPLAN contains 400+ sectors, while REMI has only 57. Second, IMPLAN initially used the 1963 regional IO data assembled in a manner similar to the national IO

data for the 50 states and Washington, DC, by Polenske (1980). When these regional technologies became more and more dated, they switched to using location quotients (LQs) to adjust the national IO coefficients, and currently they are using regional purchase coefficients (RPCs) estimated from trade flows. Starting from the initial REMI model, Treyz recognized the superiority of using RPCs over LQs.

Although adjustments using RPCs create more accurate estimates than LQs and are relatively inexpensive, some analysts, including the current authors, believe that the best method is to assemble US regional IO tables using the same data sources the Bureau of Economic Analysis uses for the national data. That is the thrust of the MITERS research, in which staff from the Massachusetts Institute of Technology and the Economic Research Service of the US Department of Agriculture are constructing multi-regional accounts at a county level and for approximately 200 sectors. They use GAMS (General Algebraic Modeling System) to create input files from census and other data to help estimate suppressed data.

Several analysts are designing new multi-regional program packages in attempts to make the models suitable for use in modern impact assessments, such as for distribution of agricultural goods, terrorist attacks, air transportation impact, and so on. Examples include Lahr who redesigned Stevens's PC-IO package, and the University of Southern California (USC) staff who designed the NIEMO model discussed above. These multi-regional models are used by numerous state and county groups, and academics.

An important extension of the direct, indirect and induced impact analysis is determination of so-called catalytic effects. For air transportation, Oxford Economic Forecasting distinguishes between supply-side and demand-side catalytic effects, with the supply-side effects indicating the performance of the economy and long-run productivity and livability, and the demand-side effects including the use of air services to transport goods, business travelers and tourists (Cooper and Smith, 2005, p. 16).

In Europe, the need to evaluate a series of large transport infrastructure projects led to IO-type new economic geography (NEG) models (Venables, 1996). Because freight and passenger transport cost reductions impact upon different sectors differently, analysts use an interregional inter-industry approach. Different regions sell varieties of the output of each sector on monopolistically competitive regional markets, linked by transport cost. Sectoral CES-aggregates of these varieties are combined in Cobb–Douglas consumption and production functions (in the latter case also with capital and labor). In these inter-industry NEG models, transport cost reductions increase each region's exports (demand) as well as imports (supply). The net economic impact may well be negative for some sectors in some regions, while causing clustering of sectors and agglomeration of economic activity in other regions (see Venables and Gasiorek, 1996; Oosterhaven and Knaap, 2003; Thissen, 2005, for seminal applications).

21.6 Conclusion

Regional analysts are incorporating new theoretical perspectives related to economic geography into their socio-economic impact models. We have reviewed the basic Types I, II and III input–output quantity and price models and summarized the state of the art in regional and interregional impact analysis. From this, we conclude that the future will continue to feature interregional inter-industry models, as the sector-specific and location-specific nature of the employment, energy and emissions impacts of all kinds of

exogenous shocks and policy measures requires such modeling. Increasingly, however, these models will feature non-linear production and consumption functions, and integrate simultaneous price and quantity impacts in models with non-perfect competition.

Obviously, the collection and processing of regional technology and regional trade data remain a time-consuming and expensive task. Given the need for regional and interregional impact analyses, in the United States several different groups are constructing multi-regional models. Because most of this work is unpublished at the present time (2008), we have presented information gleaned from websites, personal conversations and personal involvement. From this overview, we conclude that new spatial theories are needed before analysts can make significant advances on applications and that there is no real substitute to survey-based, inter-industry, interregional information and modeling.

Notes

1. The InterRegional Input–Output Software package IRIOS (Stelder et al., 2000), which is based on a flexible generalization of impact equation (21.6), is freely downloadable at www.REGroningen.nl/irios.
2. Note that de Mesnard (2006) takes offense at the use of the word 'multiplier' in this case and proposes an alternative net multiplier definition. See Oosterhaven (2007) for a reply and Dietzenbacher (2005) for an independent evaluation.
3. The first Washington State table was assembled in 1967 by Tiebout, Bourque, Thomas, and other faculty at the University of Washington, and faculty at Washington State University in Pullman, who assembled the agricultural components. Washington State tables are available from surveys conducted for 1963, 1972, 1982, 1987, 1997, and 2002 (as of 2008, the 2002 one is just being completed). Two 1997 tables were constructed, one with industrial sectors defined by the Standard Industrial Classification (SIC), the other by the North American Industrial Classification System (NAICS), with employment, income and output multipliers for each classification scheme for 38 and 62 sectors (Illman, 1996; State of Washington, 2004).
4. One example is the Port Economic Impact Kit (Klaers, Powers & Associates, 2001). It has a customized national IO model for the port's region. The direct economic impacts include those generated by the transhipment of cargo as well as capital investments made by waterfront facilities. The indirect and induced impacts are obtained from the purchases required by this direct expenditure and by the wages paid to the labourers in direct and indirect activities. Estimates are also made of the property taxes and occupation taxes paid by the various facilities.
5. The United States has 1963, 1972 and 1977 MRIO accounts for 51 regions (50 states and Washington, DC) and 79 to 120 sectors (Polenske, 1980), and will have by early 2009 1997 and 2002 MRIO accounts for 3076 counties plus about 500 ports of entry and 162 sectors; all the latter data for all years are constructed from census and other official US data, mostly using electronic files (Canning and Wang, 2005; Canning et al., 2007).
6. The website: http://www.remi.com/support/articlescomplete.shtmllist) lists an additional 23 evaluations.
7. See for details on the early IMPLAN research: http://www.fpl.fs.fed.us/documnts/fplgtr/fplgtr95.pdf.

References

Abe, K. (1986), 'Input–output tables in Japan and application for interregional analysis', presented at the 8th International Input–Output Conference, Sapporo, Japan.

Batey, P.W.J. (1985), 'Input–output models for regional demographic-economic analysis: some structural comparisons', *Environment and Planning A*, **17,** 77–93.

Batten, D.F. (1983), *Spatial Analysis of Interacting Economies*, Boston, MA: Kluwer Publishing.

Beemiller, R.M. (1990), 'Improving accuracy by combining primary data with RIMS: comment on Bourque', *International Regional Science Review*, **13** (1–2), 99–101.

Boomsma, P. and J. Oosterhaven (1992), 'A double-entry method for the construction of bi-regional input–output tables', *Journal of Regional Science*, **32** (3), 269–84.

Bourque, P.J. (1990), 'Regional multipliers: WAIO vs. RIMS', *International Regional Science Review*, **13** (1–2), 87–98.

Canning, P. and K. Polenske (forthcoming), US 1992 and 2002 Multiregional Input–Output Accounts, US Department of Agriculture, Economic Research Service.

Canning, P. and Z. Wang (2005), 'A flexible mathematical programming model to estimate interregional input–output accounts', *Journal of Regional Science*, **45** (3), 539–63.

Canning, P., A. Ismail, K.R. Polenske, S. Huang and A. Waters (2007), 'US energy consumption and food security', presented 21 October at the Associated Collegiate Schools of Planning Conference, Milwaukee, WI.
Carter, A.P. (1970), *Structural Change in the American Economy*, Cambridge, MA: Harvard University Press.
Cartwright, J.V., R.M. Beemiller, E.A. Trott and J.M. Younger (1982), 'Estimating the potential impacts of a nuclear reactor accident', Bureau of Economic Analysis, Washington, DC.
Chen, C.Y. and A. Rose (1985), 'The joint stability of input–output production and allocation coefficients', *Modeling and Simulation*, **17**, 251–5.
Chenery, H.B. (1953), 'Regional analysis', in H.B. Chenery, P.G. Clark and V.C. Vera (eds), *The Structure and Growth of the Italian Economy*, Rome: US Mutual Security Agency, pp. 97–129.
Cole, S. (1989), 'Expenditure lags in impact analysis', *Regional Studies*, **23** (2), 105–16.
Cole, S. (1997), 'Closure in Cole's reformulated Leontief model: a response to R.W. Jackson, M. Madden and H.A. Bowman', *Papers in Regional Science*, **76** (1), 29–42.
Cooper, A. and P. Smith (2005), 'The economic catalytic effects of air transport in Europe', Report prepared by Oxford Economic Forecasting, September.
Davar, E. (1989), 'Input–output and general equilibrium', *Economic Systems Research*, **1** (3), 331–44.
Davis, H.C. and E.L. Salkin (1984), 'Alternative approaches to the estimation on economic impacts resulting from supply constraints', *Annals of Regional Science*, **18**, 25–34.
de Mesnard, L. (2006), 'A critical comment on Oosterhaven–Stelder net multipliers', *Annals of Regional Science*, **41** (2), 249–71.
Dietzenbacher, E. (1997), 'In vindication of the Ghosh model: a reinterpretation as a price model', *Journal of Regional Science*, **37** (4), 629–51.
Dietzenbacher, E. (2005), 'More on multipliers', *Journal of Regional Science*, **45**, 421–6.
Eding, G.J., J. Oosterhaven, B. de Vet and H. Nijmeijer (1999), 'Constructing regional supply and use tables: Dutch experiences', in G.J.D. Hewings, M. Sonis, M. Madden and Y. Kimura (eds) *Understanding and Interpreting Economic Structure*, Berlin: Springer Verlag, 237–63.
Ghosh, A. (1958), 'Input–output approach in an allocation system', *Economica*, **25**, 58–64.
Giarratani, F. (1974), 'The effect on relative prices of air pollution abatement: a regional input–output simulation', *Modeling and Simulations*, **5**, 165–70.
Giarratani, F. (1976), 'Application of an interindustry supply model to energy issues', *Environment and Planning A*, **8**, 447–54.
Hirsch, W.Z. (1959), 'Interindustry relations of a metropolitan area', *Review of Economics and Statistics*, **41**, 360–69.
Ichimura, S. and H. Wang (2003), *Interregional Input–Output Analysis of the Chinese Economy*, Singapore: World Scientific Publishing Co.
Illman, D.L. (1996), 'A century of excellence in science and technology at the University of Washington', University of Washington, www.washington.edu/research/pathbreakers/index.html#1965.
Isard, W. (1951), 'Interregional and regional input–output analysis: a model of the space economy', *Review of Economics and Statistics*, **33**, 318–28.
Isard, W. (1960), *Methods of Regional Analysis: An Introduction to Regional Science*, Cambridge, MA: MIT Press.
Jackson, R.W., M. Madden and H.A. Bowman (1997), 'Closure in Cole's reformulated Leontief model', *Papers in Regional Science*, **76** (1), 21–8.
Klaers, Powers & Associates (2001), 'Economic impact of 2001 shipping season', Duluth Seaway Port Authority, Port of Duluth-Superior, www.duluthport.com/dsweconomicimpact2001.html.
Kop Jansen, P.S.M. and T. ten Raa (1990), 'The choice of model in the construction of input–output coefficients matrices', *International Economic Review*, **31**, 213–27.
Lahr, M. (1993), 'A review of the literature supporting the hybrid approach to constructing regional input–output models', *Economic Systems Research*, **5** (3), 277–93.
Leontief, W.W. (1951), *The Structure of the American Economy: 1919–1939*, New York: Oxford University Press.
Leontief, W.W. and A. Strout (1963), 'Multiregional input–output analysis', in T. Bama (ed.), *Structural Interdependence and Economic Development*, New York: St Martin's Press, pp. 119–50.
Madsen, B. and C. Jensen-Butler (2005), 'Spatial accounting methods and the construction of spatial social accounting matrices', *Economic Systems Research*, **17** (2), 187–210.
Miernyk, W.H. (1965), *The Elements of Input–Output Analysis*, New York: Random House.
Miller, R.E. and P.D. Blair (1985), *Input–Output Analysis: Foundations and Extensions*, Englewood Cliffs, NJ: Prentice Hall.
Miyazawa, K. (1976), *Input–Output Analysis and the Structure of the Income Distribution*, Berlin: Springer.
Miyazawa, K. and S. Masegi (1963), 'Interindustry analysis and the structure of income distribution', *Metroeconomica*, **15**, 89–103.
Moore, F.T. (1955), 'Regional economic reaction paths', *American Economic Review*, **45** (2), 133–48.
Moore, F.T. and J.M. Peterson (1955), 'Regional analysis: an interindustry model of Utah', *Review of Economics and Statistics*, **37**, 368–83.

Moses, L.N. (1955), 'The stability of interregional trading pattern and input–output analysis', *American Economic Review*, **45** (5), 803–32.
Mourouzi-Sivitanidou, R. and K.R. Polenske (1988), 'Assessing regional economic impacts with microcomputers', in R.E. Klosterman (ed.), *A Planners Review of PC Software and Technology*, Planning Advisory Service, Report Nos. 414/415, Chicago, IL: American Planning Association, pp. 101–6.
Okamoto, N. and T. Ihara (2005), 'Spatial structure and regional development in China: interregional input–output approach', Institute of Developing Economies, Japan External Trade Organization, IDE Development Perspective Series No. 5, Japan.
Oosterhaven, J. (1981a), *Interregional Input–Output Analysis and Dutch Regional Policy Problems*, Aldershot: Gower Publishing.
Oosterhaven, J. (1981b), 'Export stagnation and import price inflation in an interregional input–output model', in W. Buhr and P. Friedrich (eds), *Regional Development under Stagnation*, Baden-Baden: Nomos-Verlag, pp. 124–48.
Oosterhaven, J. (1984), 'A family of square and rectangular interregional input–output tables and models', *Regional Science and Urban Economics*, **14**, 565–82.
Oosterhaven, J. (1988), 'On the plausibility of the supply-driven input–output model', *Journal of Regional Science*, **28** (2), 203–17.
Oosterhaven, J. (1996), 'Leontief versus Ghoshian price and quantity models', *Southern Economic Journal*, **62** (3), 750–59.
Oosterhaven, J. (2000), 'Lessons from the debate on Cole's model closure', *Papers in Regional Science*, **79** (2), 233–42.
Oosterhaven, J. (2007), 'The net multiplier is a new key sector indicator: reply to De Mesnard's comment', *Annals of Regional Science*, **41** (2), 249–71.
Oosterhaven, J. and H. Folmer (1985), 'An interregional labour market model incorporating vacancy chains and social security', *Papers of the Regional Science Association*, **58**, 141–55.
Oosterhaven, J. and T. Knaap (2003), 'Spatial economic impacts of transport infrastructure investments', in A. Pearman, P. Mackie and J. Nellthorp (eds), *Transport Projects, Programmes and Policies: Evaluation Needs and Capabilities*, Aldershot: Ashgate, pp. 87–105.
Oosterhaven, J. and T.M. Stelder (2002), 'Net multipliers avoid exaggerating impacts: with a bi-regional illustration for the Dutch transportation sector', *Journal of Regional Science*, **42** (3), 533–43.
Oosterhaven, J., E.C. van der Knijff and G.J. Eding (2003), 'Estimating interregional economic impacts: an evaluation of nonsurvey, semisurvey, and full-survey methods', *Environment and Planning A*, **35** (1), 5–18.
Park, J.Y., P. Gordon, J.E. Moore II and H.W. Richardson (2007), 'Simulating the state-by-state effects of terrorist attacks on three major US ports: applying NIEMO', in H.W. Richardson, P. Gordon and J.E. Moore II (eds), *The Economic Costs and Consequences of Terrorism*, Cheltenham, UK and Northampton MA, USA: Edward Elgar, pp. 208–34.
Piispala, J. (2000), 'On regionalising input–output tables: experiences from compiling regional supply and use tables in Finland', presented at the 13th International Input–Output Conference, 21–25 August, Macerata, Italy.
Polenske, K.R. (1970), 'An empirical test of interregional input–output models: estimate of 1963 Japanese production', *American Economic Review*, **60**, 76–82.
Polenske, K.R. (1980), *The US Multiregional Input–Output Accounts and Model*, Lexington, MA: Lexington Books.
Polenske, K.R. (1995), 'Leontief's spatial economic analyses', *Structural Change and Economic Dynamics*, **6**, 309–18.
Polenske, K.R. (1997), 'Current uses of the RAS technique: a critical review', in A. Simonovits and A.E. Steenge (eds), *Prices, Growth, and Cycles*, London: Macmillan Press, pp. 58–88.
Polenske, K.R. (1999), 'Wassily W. Leontief, 1905–1999', *Economic Systems Research*, **11**, 341–8.
Polenske, K.R. and G.J.D. Hewings (2004), 'Trade and spatial economic interdependence', *Papers in Regional Science*, **83** (1), 269–89.
Polenske, K.R., K. Robinson, Y.H. Hong, X. Lin, J. Moore and B. Stedman (1992), 'Evaluation of the South Coast Air Quality Management District's methods for assessing socioeconomic impacts of district rules and regulations', prepared for the South Coast Air Quality Management District, Diamond Bar, CA.
Pyatt, G. (2001), 'Some early multiplier models and the relationship between income distribution and production structure', *Economic Systems Research*, **13** (2), 139–63.
Pyatt, G. and J.I. Round (1979), 'Accounting and fixed price multipliers in a social accounting matrix framework', *Economic Journal*, **89**, 850–73.
Pyatt, G. and E. Thorbecke (1976), *Planning Techniques for a Better Future*, Geneva: International Labour Office.
Richardson, H.W., P. Gordon and J.E. Moore II (eds) (2005), *The Economic Impacts of Terrorist Attacks*, Cheltenham, UK, and Northampton, MA, USA: Edward Elgar.

Richman, D.S. and R.K. Schwer (1993), 'A systematic comparison of the REMI and IMPLAN models: the case of southern Nevada', *Review of Regional Studies*, **23** (2), 143–161.
Schumann, J. (1968), *Input–Output Analyses*, Berlin: Springer Verlag.
Shields, D.J., S.A. Winter, G.S. Alward and K.L. Hartung (1995), 'Energy and minerals industries in national, regional, and state economies', US Department of Agriculture, Forest Products Laboratory, General Technical Report, FPL–GTR–95 Washington, DC, www.fpl.fs.fed.us/documnts/fplgtr/fplgtr 95.pdf.
Siddiqi, Y.M. and M. Salem (1995), 'Regionalization of commodity-by-industry input–output accounts: the Canadian case', Ottawa: Statistics Canada.
Spijker, J. (1985), 'Een combinatie van directe en indirecte methoden', in J. Oosterhaven and B.B.A. Drewes (eds), *Constructie en actualisering van regionale en interregionale input-output tabellen*, Arnhem: Economisch-Technologische Instituten, pp. 171–8.
State of Washington (2004), 'The 1997 Washington Input–Output Table', Office of Financial Management, State of Washington, www.ofm.wa.gov/economy/io.
Stelder, T.M., J. Oosterhaven and G.J. Eding (2000), 'Interregional input–output software', IRIOS 1.0 Manual, University of Groningen, www.REGroningen.nl.
Stevens, B.H. and G.I. Treyz (1986), 'A multiregional model forecast for the United States through 1995', *American Economic Review*, **76** (2) (May), 304–7.
Strout, A.M. (1967), 'Technological change and the United States energy consumption, 1939–1954', PhD Dissertation, University of Chicago.
ten Raa, T. and J.M. Rueda-Cantuche (2003), 'The construction of input–output coefficients matrices in an axiomatic context: some further considerations', *Economic Systems Research*, **15** (4), 441–55.
Thissen, M. (2005), 'RAEM: regional applied general equilibrium model for the Netherlands', in F. van Oort, M. Thissen and L. van Wissen (eds), *A Survey of Spatial Economic Planning Models in the Netherlands*, Rotterdam: Ruimtelijk Planbureau/NAi-uitgevers, pp. 64–86.
Treyz, G.I., A.F. Friedlaender and B.H. Stevens (1980), 'The employment sector of a regional policy simulation model', *Review of Economics and Statistics*, **62** (1), 63–73.
Vaccara, B.N. (1969), 'Changes over time in input–output coefficients for the United States', in A.P. Carter and A. Brody (eds), *Applications of Input–Output Analysis*, Amsterdam: North-Holland, pp. 238–60.
van Dijk, J. and J. Oosterhaven (1986), 'Regional impacts of migrants' expenditures: an input–output/vacancy-chain approach', in P.W.J. Batey and M. Madden (eds), *Integrated Analysis of Regional Systems*, London: Pion, pp. 122–47.
Venables, A.J. (1996), 'Equilibrium locations of vertically linked industries', *International Economic Review*, **37** (2), 341–59.
Venables, A.J. and M. Gasiorek (1996), 'Evaluating regional infrastructure: a computable equilibrium approach', mimeo, London School of Economics.
West, G.R. (1990), 'Regional trade estimation: a hybrid approach', *International Regional Science Review*, **13** (1–2), 103–18.

PART V

REGIONAL GROWTH AND DEVELOPMENT POLICIES

22 Institutions and regional development

T.R. Lakshmanan and Ken J. Button

22.1 Introduction

Institutions have long been recognized as important in shaping economic development, but until comparatively recently the amount of analytical analysis refining the exact linkages has been comparatively limited. There had been a tendency in the past for political scientists, sociologists, economists and geographers, such as Veblen, Kenneth Galbraith and Fagg Foster to describe institutions and critique conventional analysis that did not adequately embrace them. The peculiarities of institutional structures, their dynamics, and roles at the regional level tended to be treated either implicitly in more generic analysis or as particular, and special, features of individual regional case studies. Indeed, the definition of institutions and boundaries to institutions were ill-defined. This has changed, with institutional analysis within disciplines such as economics, and most notably the new institutional economics as developed by Coase, Oliver Williamson and Matthews, embracing the need to place markets within their institutional context and within a more complete analytical context.[1]

Institutions are now seen as comprising a set of formal and informal rules, including the conditions of their enforcement. Douglass North (1994) views institutions as made up of formal constraints (rules, laws and constitutions), informal constraints (conventions, norms and self-imposed codes of conduct) and their mechanisms of enforcement. By defining and delimiting the set of options available to individuals, institutions can reduce uncertainty, simplify action choices, and offer an incentive structure for activities and interactions among economic agents.[2] Institutions not only define the range of actions available to individuals; they are also shaped by individuals and render individual interaction possible. Institutions thus function as a rational framework or infrastructure for governing interactions among individuals. In this sense, institutions function as public goods, available to and shared by many, and are non-excludable (Nelson, 1959).

Organizations are defined as institutions together with the people taking advantage of them (North, 1990), or 'the personal side of the institution' (Schmoller; quoted in Furubotn and Richter, 1997). In some parts of neoclassical economic theory, organizations do not appear, so that institutions become 'the grin without the cat, the rules of the game without the players' (Furubotn and Richter, 1997).

Here we focus on the different institutional mechanisms that allow for the coordination of regional economic activities in modern capitalistic economies. The chapter inquires into the logic and functions of economic institutions. We also trace the emergence of regional institutions in the light of the coherence of the institutional attributes with the structural conditions prevailing in a regional economy and examine the subsequent development and persistence of institutions.

Economic institutions are not viewed, however, as static entities but are responsive and evolve under changing conditions. This issue is examined in the context of rapid technical change and market integration across vast distances in the contemporary, dynamic

regional economies of the world. Such technical and market evolutions set in motion changes in the nature and patterns of economic interactions, and broader economy-wide incentive structures, thereby creating an environment for the rise and evolution of new institutional forms. In knowledge-oriented regional economies, innovation-intensive networks emerge as appropriate economic governance mechanisms. They emerge as the outcome of complex evolutionary mechanisms in response to the requirements of the growth and development of regional economies at different times and in different places. Different institutions appear in a variety of combinations and complement one another in the performance of the regional economy.

The second section of the chapter offers a brief tour of the various concepts in economic theory pertaining to institutional origins, diversity and evolution. It initially looks at the contributions of economic theory to the understanding of institutions,[3] followed by discussions of the objectives and functional logics of three major institutions that have guided economic coordination over the last two centuries or so: markets, private hierarchies and the state. The kinds of changes in the structure of the broader regional economy that alter the nature of economic incentives are identified and new incentives that may require new functional logics in the institutions that govern economic transactions are reviewed. In particular, we describe the changing structural conditions in dynamic regional economies and how these have supported the emergence of new institutional forms.

Section 22.3 deals with the emergence of the 'knowledge economy' and the market expansion to global scale in dynamic regional economies around the world, and the major changes these developments pose for the nature and patterns of interactions among regional economic agents, and the broader economy-wide incentive structures.

In contemporaneous regional economies, firms often adopt new institutional forms of networks that involve horizontal relationships with other regional economic agents, creating a hybrid of the two broad organizational principles of economic institutions – self-interest and social rules – with the objective of addressing some of the limitations of classical markets (for example an inability to capture innovation spillovers) and large firms (that are slow to respond to technological change).

In section 22.4 some broader perspectives on regional economic governance are offered. First, while recognizing that each economic institution (the market, firm, state, networks, associations, and so on) has strengths and weaknesses, the preferable approach is not to favor one institution but to combine them according to objectives, resources and the attributes of the goods and services. Second, we broaden the theoretical perspective in recognizing that the operation of regional economic institutions are constrained by the social context in which they are embedded (Polanyi, 1957 [1944]; Granovetter, 1985). Depending on the nature of embeddedness, some of the resulting forms of economic governance display more of the social rule-based organizing principle, as compared to the self-interest principle.

The chapter concludes with an overall framework of organizing the different economic institutions – one that relates the underlying institutional organizing principles to the nature and pattern of organizing economic actors within an institution.

22.2 Institutions: emergence and evolution

Historically, uncertainty in human interactions has been reduced or absorbed by various kinds of positional or status systems, by relationships such as those operating in a family

or clan. Such dependence relationships and exchanges with known others, as well as with those that one shares ideals and norms, create a common understanding and trust as well as reducing opportunistic behavior and 'moral hazard', and are crucial for economic interactions. Others associated with calculative action have, however, increasingly overtaken such forms of human interactions. In this mode individuals attempt to lower uncertainty by way of calculative relations – as specified by self-interest or *homo economicus* rationality. This is why tribes were formed and towns built. Understanding the theoretical concepts underlying the origins, functions and evolution of economic institutions as developed in economic theory, the manner in which structural changes in the broader economy are derived from technological change and market widening, and new perspectives deriving from the increasing recognition of the 'embeddedness' of economic institutions in the broader society and culture, offer a basis for our discussions.

Institutions in classical and neoclassical economic theory

In classical economics, institutions played a role, as evidenced in the writings of Adam Smith and J.S. Mill (1857) and later in the work of Alfred Marshall (1920), who noted that the institutional structure influences economic behavior. However, as neoclassical economic theory took hold, institutions were often viewed as exogenous phenomena and marginalized theoretically. Given an exogenous system of rules and norms, the neoclassical focus was one of optimal combinations of land, capital and labor as derived from utility- and profit-maximizing behavior of economic agents. In this theory, the price system is the only explicitly modeled process used to coordinate different economic activities. In a market-driven system, administrative coordination is viewed as unnecessary – with the coordination inside even institutions, such as firms, viewed as occurring in a 'black box'.[4]

The emergence of institutions in this framework is explained in two ways. At one end is the 'spontaneous' origin of institutions 'from below' through the operation of individual self-interest that Friedrich Hayek (1973) describes as deriving from 'evolutionary rationalism'. At the other end, institutions may result 'from above' through deliberate design or 'made or grown to order', by the actions of some authority (parliament, an entrepreneur, a group, and so on). The respective situations are viewed as spontaneous and intentional governance, and examples are, respectively, a market community and a firm.

Resources are needed to operate this exogenous system of institutions and assure adherence to the rules and norms. Understanding this leads to the abandonment of the neoclassical assumption of a 'frictionless' economic system and the recognition of transaction costs.[5] In a world of frictional costs, property or contractual rights cannot be instantly defined, monitored, enforced or transferred without an outlay of resources. In other words, such transaction costs are the search and bargaining costs of using the market, as well as the costs of administration in a private hierarchy such as a firm. Indeed, these transaction costs appear extensively and significantly in modern market economies – estimated in some cases at 50 per cent to 60 per cent of net national product (Furubotn and Richter, 1997).

In the history of economic thought, reaction against the exclusion of institutions and organizations and transaction costs in neoclassical theory has appeared in two broad streams. The first stream is represented by the 'old institutional economics' or the

institutional school of American economists (for example Veblen, Commons, Mitchell and, more recently, John Kenneth Galbraith), and the German Historical School (Gustav von Schmoller). John R. Commons (1934) suggested that the collective control of individual transactions should be the focus of institutional economics, arguing that the 'conflict of interests' is predominant in transactions among economic agents, rather than 'harmony' as assumed in classical and neoclassical theory. He defined an institution as collective action governing individual action. Cooperation, in such a context, derives from deliberate action intended to create a new harmony of interests among the cooperators – not a result of 'harmony of interests' as assumed in classical and neoclassical theory.[6]

New institutional economics

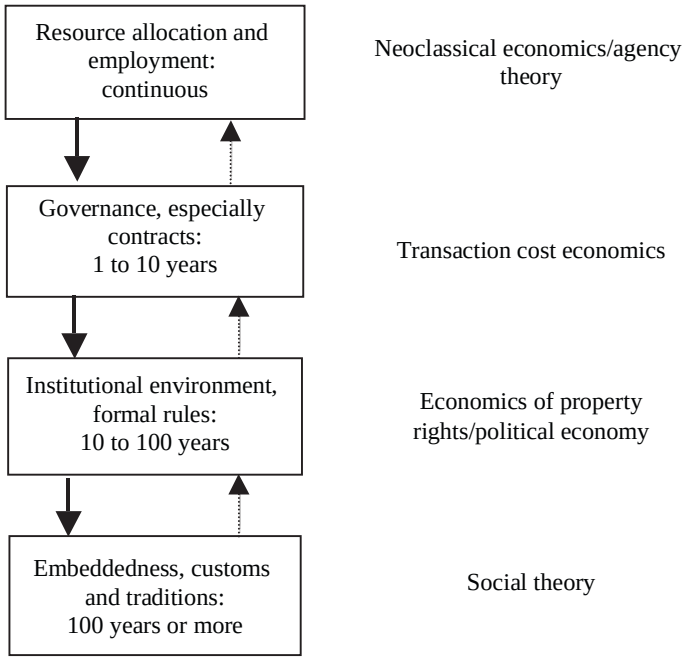
The second school of theoretical inquiry is the new institutional economics (NIE), associated with the work of Ronald Coase, Oliver Williamson, Douglass North, Alfred Chandler and others. The NIE theorists reject the two central ideas of neoclassical theory: costless transactions and neutral institutions.[7] They accept the notion of positive transaction costs and the view of institutions as influencing transaction costs and individual incentives and thus economic behavior. Because of transaction costs, the neoclassical view of perfect rationality has to be replaced by some version of 'bounded rationality'.[8] Further, these theorists elaborate on the concept of transaction costs and related notions and view institutions as endogenous variables in the economic model.

Within this framework, neoclassical economics and conventional neoclassical views of markets are set within a larger framework (Figure 22.1) that considers not only the immediacy of markets but also the institutional context, both formal and informal, in which they function. The figure also offers some very broad indication of the degree of rigidity, in terms of years it takes for change to occur, associated with the various elements.

The consequent theoretical work of the NIE derives from subfields such as transaction cost economics, property-rights analysis, contract theory, comparative systems, and so on. The resulting literature is rather diverse in content, method and style – some verbal and discursive, others formal and mathematical (for example the work of information and contract theorists like Joseph Stiglitz (1985) and Bengt Holmstrom). This review is of the first type.

Oliver Williamson's (1985) definition of a transaction is when 'a good or service is transferred across a technologically separable surface – as one stage of activity terminates and another begins'.[9] Transactions occur among a variety of economic actors – researchers, capitalists, manufacturers, labor, suppliers of raw materials and intermediate products, and others who jointly address problems such as raising capital, deciding on product standards, setting levels of output, wages and prices, and marketing products to consumers. John R. Commons (1934) views transactions as also involving the transfer of sanctioned property rights between individuals

Economic and political costs are involved in these transactions. The former pertain to market transaction costs, and the costs of the authority to issue orders inside a firm (managerial transaction costs). The use of the market involves deciding on who to negotiate and sign a contract with, and later monitoring contract performance. Coase (1960) recognizes three types of such costs: search and information costs (costs of contract preparation); bargaining and decision-making costs (contract conclusion costs); and contract monitoring and enforcement costs. Managerial transaction costs involve the costs of setting up, maintaining and changing a firm, and the costs of running it.



Source: Williamson (2000).

Figure 22.1 Characterization of where institutions become important

Political transaction costs arise because the economic, market and managerial transactions can occur only in the context of political institutional arrangements consistent with a capitalist market environment. That implies the need for the existence of a regional or national organization and the associated public goods and costs. Such costs of supplying goods by collective action are the political transaction costs. They comprise of the costs of setting up, maintaining and changing the background formal and informal political organization (legal framework, administration, judiciary, and so on) and the costs of running the polity (current outlays for measuring, monitoring and enforcing compliance).

These transaction costs have to be incurred in order to facilitate orderly transfers of property rights in economic transactions. Broadly stated, such costs can be viewed as the costs of specialization and division of labor. Further, their level can vary with the nature of the broader society and the constituent individual behavior. If there were greater trust and shared norms or mental models among individuals in a region or nation, monitoring and enforcement costs can be lower – suggesting interrelationships between trust, social values and the institutional framework in a region or nation.

In addition to the concept of transaction costs associated with the exchange process, the NIE theorists utilize two other analytical constructs to support their insights on the origin and evolution of economic institutions in a region or nation. These pertain to the analysis of property rights and economic theory of contracts.

The right of ownership of an asset consists of the right to use it, change its form and substance, and transfer some or all rights as desired. The assumption is that property

rights are assigned according to the principle of private ownership (as in a liberal state), giving the individual discretionary power over resources and providing a basis for competitive markets. In a world with positive and ubiquitous transaction costs, uncertainty and asymmetric information prevail. Under these circumstances, the pattern of property rights in an economy affects economic incentives and thus affects behavior and economic outcomes. This allows NIE analysts to analyze the impact of property-rights arrangements on regional or national economic outcomes.

The third area of NIE concern is the economic theory of contracts. The NIE theorists formally approach transaction costs, or the information costs of economic interactions, by analyzing their effects, starting with situations when costs can be rather high. These are situations that are either too costly or impossible to find because they involve imperfect foresight about future conditions and asymmetric information when other agents know more or better than others.

Imperfect foresight prevents the preparation of a contract that embraces all conceivable contingencies in the future. Risk-averse employers and employees, for example, may cover the resulting risks through labor contracts – with firms offering workers a lower average wage level but insurance against variations in real wages. Another consequence of imperfect foresight is the difficulty, or near impossibility, of contract enforcement, leaving gaps in contracts and room for later adjustment. Such issues are addressed by incomplete contract theory. Finally, there are informal agreements between a firm and its customers – the theory of self-enforcing agreements shows that honesty is the best policy for self-interested economic agents.

Asymmetric information typically occurs in the principal–agent problem when an agent (for example firm manager or employee) enjoys an informational advantage over the principal (firm owner or employer). Such occurrence is observed in two contexts: precontractual asymmetric information (adverse selection), and post-contractual asymmetrical information (moral hazard) – with the better-informed party potentially engaging in opportunism.[10]

Basic institutions of economic governance

The overall insight is that several economic coordinating mechanisms or institutions can emerge in economies. The emergence of these various institutional types depends on the circumstances prevailing in an economy and if these circumstances are consistent with the inherent logic of a particular institutional mechanism. Further, there are arguments that the different institutions of capitalism complement one another and cannot generally function in isolation from one another (Hollingsworth and Boyer, 1997; Baumol et al., 2007).

Markets In very many countries there is, particularly with the recent advent of neoliberal ideas, a widespread view that the most efficient institution for coordination of economic activity is the 'market'. The neoclassical framework underlying this view is rationalistic, individualistic and utilitarian. In the traditional version, market coordination occurs when economic actors remain autonomous, pursue their interests vigorously and engage in decentralized arm's-length bargaining. These economic actors indicate their preferences and prices through relatively comprehensive contracts, which are self-liquidating with completion of contract performance – without requiring further

interactions among economic actors (Williamson, 1985). The objective is to complete on-the-spot instantaneous transactions without any implications for future activities or strategies. This is clearly a stylized characterization of the classical market – the pure competitive spot markets – however, one can recognize a large number of variants and modifications of this form around the world.

Economists in this tradition hold an adverse view of other institutional mechanisms for economic coordination. In their view, most alternative institutional forms and the state do more harm than good. Such perspectives are often supported by references to the experience in recent times with Keynesian interventions, which many argue contributed to stagflation in the 1970s, and the economic collapse of the planned economies in Eastern Europe and elsewhere in the late 1980s. As a consequence, this emphasis on market institutions has gained currency not only in economic thinking, but also in scholarship in related areas such as public choice theory (Coleman, 1990), political science (Ostrom, 1986) and law (Posner, 1977).

In spite of this the market should not be viewed as the ideal and universal institutional arrangement for coordinating economic activities. When efforts are made to organize economic activities exclusively in terms of markets, they can function well and provide social optimum only if some structural conditions obtain in the larger economy (see Figure 22.1). Such structural conditions include: absence of uncertainty, reversibility of transactions, completely contingent markets, non-occurrence of increasing returns to scale, and no collusion among economic actors. Adam Smith's invisible hand solution applies under these circumstances with an efficient and harmonious outcome. More generally, however, we can state that the organization of the economy solely in terms of markets can generate several types of inefficiencies and drawbacks in economic coordination. Each of these drawbacks offers fertile ground for institutional reform and the emergence of a new institutional form. The analytical challenge is to understand and identify the specific institutional forms that offer viable economic performance under different conditions.

Private hierarchies Some economists with an NIE theoretical persuasion have highlighted the first deficiency. They argue that private hierarchical organizations such as firms are more efficient under certain conditions than markets and contracts (Williamson, 1975, 1985; Coase, 1988). When economic actors carry out their transactions within a firm or a private hierarchy they can reduce transaction costs, enhance efficiency and reduce the opportunism inherent in exchange relationships. The incentives to realize lower production costs and greater scale economies promote economic agents to act within firms (hierarchies) rather than outside the firm (that is, the market). This process of internalizing economic transactions inside a firm lowers risk and uncertainty and confers the capability of achieving lower costs and higher levels of productivity, in the context of large markets and capital-intensive but fairly stable technologies. Alfred Chandler (1977) noted the emergence of the new institutional form of a large firm or corporation, that is, a vertical or hierarchical organization of economic actors in the private sector that appeared in the US in the nineteenth century, first in the railroad industry and later in many manufacturing sectors.[11]

The state While markets exhibit strengths in the efficient provision of private and divisible goods, they evidence weaknesses and require supplementary support in two broad areas.

First, state intervention is generally required as a minimum to enforce the rules among the interacting economic agents in markets and among firms, as the market and managerial transactions occurring among economic actors in the market and within firms can be efficient only in the context of political institutional arrangements that are consistent with a capitalistic market order. Such institutional arrangements pertain to the supply of public goods for collective action. Such public goods relate to the creation, maintenance and change of a background formal and informal political organization for structuring economic transactions – the legal, judicial and administrative frameworks for facilitating economic transactions, and the process of running such a polity. The provision of such organizations and the associated public goods is the province of the state as an institution.[12] The state sanctions and regulates other non-state coordinating mechanisms, and is the ultimate enforcer of the rules of other mechanisms that enumerate and enforce property rights and set monetary and fiscal policy. By offering such an organizational apparatus related to contract enforcement, property rights transfer, and so on, the state facilitates economic transactions in the market and among firms – thus promoting specialization and division of labor in the larger economy. This key role of the state as another institution of economic coordination is acknowledged from even the minimalistic perspectives of economic theoretical devotees of the market; indeed Adam Smith acknowledges it.[13]

The state in practice comprises of several institutional arrangements such as departments, regulatory agencies, and so on, which in turn have different objectives, processes and consequences. In addition, the state delegates some of its functions to private interest groups. When this happens, as in the delegation of regulation and monitoring of stock exchanges, professions and occupations, an associational form of governance occurs (Coleman, 1990). It is worthy of note that the state has sometimes been an economic actor engaged directly in exchange and production activities. Globally, this role is declining sharply with the collapse of the Socialist bloc and the rise of neoliberal ideologies, but still continues even in oft-perceived 'market' economies such as the US where such things as roads, airports, seaports and other pieces of major infrastructure are within the direct domain of national, state or local government.

The second argument for the state as an economic institution derives from the notion of market failure when markets result in problems associated with public goods, spillover effects, indivisibilities, increasing returns to scale, and so on.

During the 1930s and 1940s, a variety of perceived market failures led to state intervention in the form of extensive regulation and monitoring and guiding of macroeconomic activity. Recent theoretical developments suggest that modern industrialized economies will be unable to realize their full potential for economic growth and development without market interventions. There have been demonstrations, for example, that investments in education have significant spillover effects on innovation and quality of health, thereby enhancing economic productivity and growth. Given that these benefits are partly external to those educated, the level of investment in education will be below the optimal for the regional or national economy. Similarly, individuals are likely to underinvest in prevention and healthcare in the context of inadequate insurance or welfare. In the context of the spillover effects of education, growth in knowledge and some types of infrastructure, the 'new endogenous growth theory' argues that certain supply-oriented state intervention can promote market performance and long-run growth (Romer, 1986, 1990, 1994).

Table 22.1 Differentiating characteristics of basic economic institutions

Institutional characteristics	Market	Private hierarchies	State
A. Organizational attributes	Free entry and exit Voluntary spot exchanges	Bureaucratic rules and trends Exchange based on asymmetric power	Public hierarchy unilateral action
B. Compliance mechanisms	Legal enforcement Private property	Institutional rules Corporate culture	Coercion Social rules and norms
C. Institutional failure			
Enforcement	Collusion Imperfect competition	Opportunistic behavior	'Public Failure' Lobbies capturing public-interest objectives
Public goods and externalities	Inability to produce public goods Difficulties in adapting to technical change and innovation	Slow reaction to technological change	Can provide public goods, but in what quantities? Difficulty in inducing technical change
Efficiency	Some basic social relations not producible by pure market	Difficulties in adapting to cooperative behavior	Bureaucratic ambience Higher cost of services
Equity	Promotion of income and wealth inequalities	Management overload and inequality	Power and privilege may enhance inequality

Source: Based on Hollingsworth and Boyer (1997) and Furubotn and Richter (1997).

When negative externalities occur because of an institutional inability to allocate property rights adequately, there is the case for the state as an economic coordination mechanism. If clean air or water or urban road space are viewed as free goods, households and firms have little incentive for preserving the environment or the quality of urban space. By creating environmental standards and mechanisms for allocating polluting rights or urban road space rights (a quasi-market mechanism), the state institutions support more efficient economic transactions.[14]

Table 22.1 displays the differentiating attributes of the three basic institutions – markets, firms and the state – widely used for economic coordination around the world. It highlights some of the strengths and weaknesses of each institutional mechanism. The institutions are differentiated by organizational characteristics, compliance mechanisms and type of institutional failure.

The institution of a market displays voluntary exchange and free entry and exit attributes, operates under an ideology of market legitimacy and legal enforcement by the state, and provides static efficiency, especially in the provision of private and divisible goods. However, its weaknesses include the poor ability to produce a variety of quasi-public goods such as education, innovation, transport and other infrastructure, environmental quality, and so on. Further, in its purer forms, the market promotes inequalities.

The institution of the private hierarchy is a vertical organization of economic agents, in this case in a firm, where exchange is based on asymmetric power and bureaucratic ambience, and offers efficiency by lowering transaction costs, minimizing opportunism and exploiting economies of scale (Coase, 1937). However, weaknesses of private hierarchies can include: inertia, especially in non-competitive markets, involving slow reaction and adjustment times to technological change; the possibility that governance costs of management may overwhelm the benefits of the firm's internal division of labor; and the adverse effects on economic inequality.

The institution of the state is organized by coercion and power relations, and ideally provides the minimal legal enforcement required in the operation of the market and the firm. The state mechanism, while solving market failure, can provide public goods – though it has difficulty in offering them in precise quantities. However, there can be 'public' or 'intervention' failure in the sense that bureaucrats with delegated power can substitute their own objectives and render public actions inefficient. Dictatorships are perhaps the obvious extreme manifestation of such capture of a system.

Institutional dynamics and evolution

Douglass North (1990, 1994) views institutional change as mainly triggered by changes in relative prices or in individual preferences. These prices embrace costs of market transactions, the costs of acquisition of technologies or competences, and relative factor prices. Such price changes flow from activities of the innovation creator and the entrepreneur, who shape the nature and direction of competences that economic actors now need. North suggests that adaptive change on the part of the economic actor arises in this context and activates the mechanisms of institutional change.

Thus existing institutions are buffeted, shaped and reshaped by historical and emerging changes in technology and markets and the consequent evolving nature of economic interactions in society. Such emerging changes in economic interaction patterns represent and simultaneously provide arenas for experiments, trial-and-error processes, and learn-

ing among economic agents at the micro level and potentially at the macro institutional level. Sven-Erik Sjostrand (1995) suggests that the gap or mismatch in the incentives at the micro and macro institutional levels is the driving force behind institutional change. Institutions are thus viewed as mechanisms 'that coordinate, regulate, and stabilize human activities at a macro level, while simultaneously functioning as part of the raw material of change on the micro level'.

The rise of the knowledge economy in recent years, and its predominant spatial location in a number of dynamic regional economies in North America, Europe and Asia, offers such a context of major technical and market evolution and the consequent modification in the nature and structure of incentives for economic actors – a fertile ground for the reform and evolution of new institutions.

22.3 Dynamic regional economies and the rise of networks

Regions are territorial units with incomplete or no political sovereignty within their borders. They exhibit over time different patterns of economic specialization, growth and development, and production systems. In regions where the product demand is stable and undifferentiated, and the technology standardized and changing infrequently, firms generally organize production in large vertically integrated units that exploit economies of scale and density by using single-purpose machines and low-skilled workers in fashioning standardized products (Chandler, 1977). The automobile and textile industrial districts exemplify this type of system. A second class of regional economies is characterized by differentiated product demand, volatile markets and rapid technical change; firms choose innovative and flexible production strategies – knowledge-intensive, cooperative–competitive hybrid approaches, flexible machines, labor, and so on. In such dynamic or creative and cooperative regions, innovation and lowering of the consequent adaptive costs become the critical competitive factors for firms.[15]

A key aspect of recent technical change is the maturation of the knowledge economy and the arrival of knowledge-rich technologies in various aspects of production – design, fabrication, input and output logistics, marketing, after-sales services, and so on. In this increasingly knowledge-intensive context, value derives from knowledge, and enterprises seek to add value to their core competencies by taking advantage of complementary assets and capabilities of other enterprises. At the regional level Doloreux and Parto (2005) highlight the roles of knowledge transfer and knowledge infrastructure in complementing the strategies and performance of firms in stimulating innovative activities.

This development introduces new competitive factors and alters the incentive structures in regional economies typically populated by dynamic small and medium-sized enterprises (Alter and Hage, 1993). First, innovation becomes a more pivotal competitive factor than cost reduction and productivity enhancement, emphasized by NIE theorists such as Williamson (1985) and Chandler (1977). Second, the firms become concerned with the reduction of a new class of costs that they confront in this knowledge economy. These are adaptive costs incurred by the firm as it monitors the environment for changes in technology and products, identifies competitive strategies, and implements such strategies quickly enough to retain or improve market share (Alter and Hage, 1993).

This combination of the criticality of innovator and entrepreneur and adaptive cost reduction shapes the nature and structure of incentives for regional economic actors. It requires new competences on the part of economic actors, and new patterns of

Table 22.2 Essential characteristics of networks

Characteristic	Description
1. Organizational attribute	Voluntary interchanges over time periods Multilateral exchange Semiformal membership
2. Compliance mechanism	Resource interdependence Contractual bonds Trust developed among agents outside economic arena
3. Institutional Failures	
Enforcement	Need for an external enforcement authority
Public Good & Externality	Strong in the provision of enhanced quality of goods and training
Efficiency	Efficiency in industry with complex and rapidly changing technology
Equity	When widely developed into industrial districts, networks can promote greater equality, if weakly developed, networks can enhance social inequality

interaction among them to support the creation and maintenance of new knowledge and its commercial application. North (1990) suggests that adaptive change on the part of the economic actor arises in this context and activates the mechanisms of institutional change.

Spurred by the need to be innovative and to lower adaptive costs, firms form alliances and joint ventures to access competences, share R&D costs, and to be agile in the face of competition. Participants can remain small (or become small through downsizing) in this cooperative framework, and realize the benefits from the joint products, while lowering the adaptive costs. In other words, the network participants are functionally interdependent but autonomous. Clusters become important (Enright, 1998).

In this context of an innovative region, being small for a firm has advantages, as contrasted with large firms that flourish when confronting only transaction costs emphasized by NIE. Further, small firms staying in these networks learn over time the tacit knowledge of the network partners and build trust (Polyani, 1983 [1967]).[16] As tacit knowledge gained from this cooperation among partners grows, the firm's most valuable asset – its human capital – also increases. The growth in tacit knowledge can pave the way for more innovative products and services. Further, other developments are steering firms away from autonomy. For example, small high-tech firms sensing new or niche markets and lacking capacities for large-scale production and distribution may form joint ventures to access such expertise. Thus the logic of competing over knowledge and adaptive cost reduction leads to more complex coordination problems than those implied by the NIE logic of transaction cost reduction and productivity enhancement.

The institutional mechanism that coordinates such complex cooperative relationships among economic actors is the network (Table 22.2). While networks exhibit some features of self-interest and some of social obligation, they are not a halfway point between a firm

and a market (Alter and Hage, 1993). Networks represent forms developed in response to the changing context of desired economic interactions among economic agents in innovative regions. Networks differ from the hierarchical coordination of the firm since each participant has autonomy, because networks have 'visible hands' in the form of complex decision-making groups at multiple levels. Networks comprise sometimes of only firms, at other times of firms, public sector actors and social sector actors.

A regional economy of networks becomes increasingly a system of cooperative interactions among economic actors or a web of links between individuals, firms and organizations, with links based on knowledge assets and evolving through cooperative learning processes.[17]

22.4 Perspectives on modern regional economic governance

Multi-institutional view of regional economies

We have seen that until recently, the problem of economic coordination was largely focused, particularly in the Anglo-Saxon world, in economic analysis of the institution of markets. In recent decades, however, many economists have observed that markets for commodities such as labor, credit or natural environment operate differently from those for typical manufactured goods, because of issues related to fairness and moral hazard (Coase, 1988). Further, in the context of problems such as providing public goods, spillover effects, indivisibilities and incomplete information, the restoration of minimal efficiency in the market requires state intervention. Hence, markets are embedded in a nexus of state interventions and obligational rules, without which the market does not function well. Further, some sources for competitiveness exist at the regional level, where economic actors enter into horizontal relationships called networks that nurture trust and tacit knowledge, thereby delivering individual and collective results that overcome some deficiencies of private hierarchies and of markets.

In other words, one needs to take a multi-institutional view of regional economies of contemporary capitalism. There is no single optimal institutional arrangement for organizing economic activities, but a range of institutions of economic governance conditions (Boyer, 1997). Each institution exhibits a set of objectives, functions, strengths and deficiencies. They cannot generally operate in isolation from one another, but complement one another. The challenge for a region is to identify the institutional forms that deliver viable economic performance under the prevalent technical and social conditions – leading to the coexistence of alternative institutions in a regional economy. Clearly, the combination of economic institutions in a region depends upon the consistency between the objectives and functions of different institutions and the characteristics of the goods and services in that regional technical/market context.

The social embeddedness of institutions

Institutional economists recognize that norms differ between societies, for example the Muslim approach to interest rates is fundamentally different to the Christian attitude, and Communist regimes take a different approach to personal ownership than do capitalist systems. These are embedded social norms that underlie the basic ways societies function, and determine the mechanisms that allocate resources between regions and, *ipso facto*, the speed and nature of their economic development (see Figure 22.1). These embedded

social attitudes do change, but infrequently, and often as the result of major military or economic failures of the old set of institutions.

Thus while the shorter-term economic analysis of institutions is typically decontextualized, anthropologists and sociologists suggest that economic institutions are constrained by the social context in which they are embedded (Polanyi, 1957 [1944]; Granovetter, 1985; Sabel, 1997). The trend in the twenty-first century seems to be for greater coordination in systems with more engagement between the public and private sectors and a move away from the statist approach of the Soviet states and the more laissez-faire norms of Fordism. This is particularly so regarding technology development and transfer.

In the US, multilateral relations involving many private corporations, university research facilities and US government departments made possible the emergence of new technologies and products (computers, semiconductors, integrated circuits) ahead of commercial markets (Nelson, 1982). Other forms of promotional networking include: the German system of apprenticeship training that involves the collective efforts of unions, business associations, students, schools and the state; and the cooperative activities of different firms in the Jutland, Denmark and north-eastern Italy (the so-called *terza Italia*) (Hollingsworth and Boyer, 1997). At the other extreme there are supranational relations, where combinations of nation states may be viewed as 'regions' that have combined to engage in ultra-costly research into longer-term science – for example, the CERN project involving particle physics or the joint initiatives into the development of the space economy.

Another form of multilateral coordinating institution typically appearing at the national level is the delegation of coordination by the state to private groups in capitalistic economies. Associations coordinate economic actors engaged in the same or similar sectors – as contrasted with markets, private hierarchies or networks which coordinate activities among different types of actors (producers with customers or suppliers, capital with labor, and so on). Typical examples of associations are labor unions and business associations.

The 'architectures' of regional economic institutions

As one way of illustrating the relationships between the different regional economic institutions in a framework that embodies the key insights of contemporary institutional theory, we build on an approach developed by Hollingsworth and Boyer (1997). This approach relates the different regional institutions in terms of two underlying concepts: the forces that motivate economic actors, and the framework within which economic interactions are organized (Figure 22.2).

Conventional neoclassical economic theory asserts that self-interest motivates the economic agent, and under some structural conditions the market functions well through the invisible hand theorem.[18] Alternatively, other social sciences such as sociology, anthropology and political science, and some branches of economics, most notably behavioral economists, often emphasize the motivating principle of obligation and compliance with social rules as underlying human actions. We distinguish here the different regional institutions along this motivational dimension in Figure 22.2. A second horizontal dimension along which the institutions can be arrayed is the manner in which the interactions among economic actors are organized. At one end is the horizontal coordination among many and relatively equal economic agents, for example a market. Markets (cell A) thus

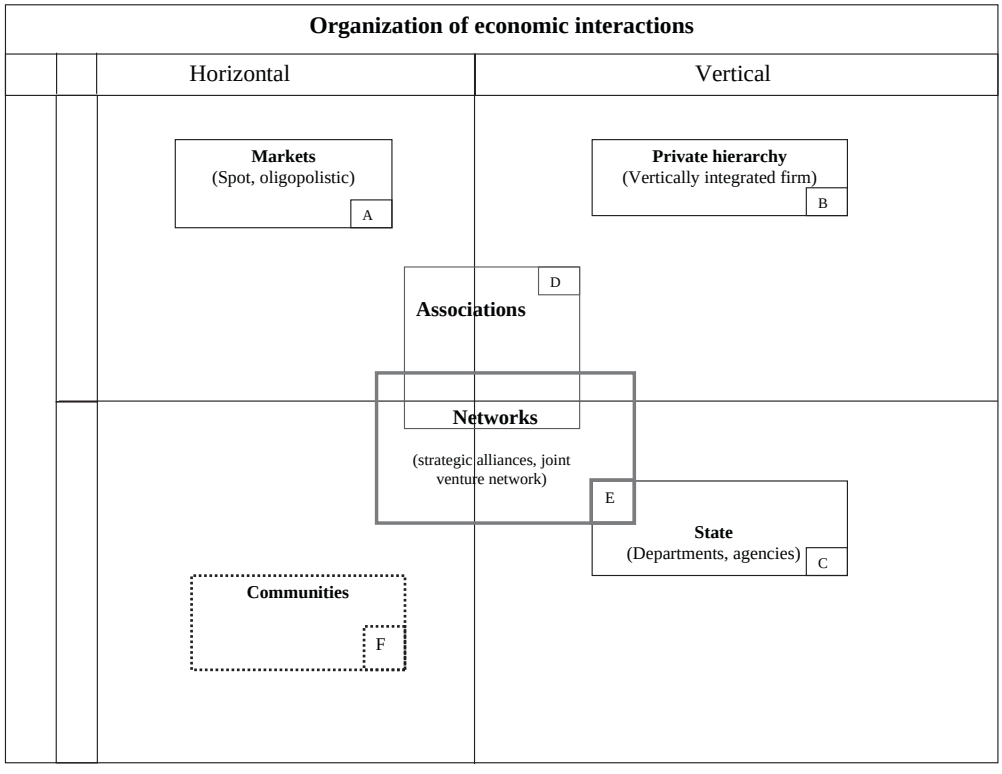


Figure 22.2 A comparison of contemporary economic institutions

blend self-interest with a horizontal organization of economic interactions that yields an *ex post* equilibrium. At the other end of the continuum, the economic agents are organized in a hierarchical firm with unequal power relationships. The actors are 'hierarchically' integrated into a firm (cell B) with its strengths and weaknesses as an institution. The institution of the state (cell C) is also hierarchically organized but motivated by an obligational or social value framework.

In the lower left part of the figure, the motivating force underlying economic interactions is a greater orientation towards collective behavior – as guided by social rules – in order to address issues of common interest. An example (cell F) is the institution of communities, where the coordination is based on non-economic criteria such as reciprocity, obligation and trust.[19] The institutions of networks (cell E) and associations (cell D) blend more centrally the self-interest and obligational principles, and the horizontal and vertical patterns of interactions among economic agents.

22.5 Conclusions

Institutions are numerous and vary considerably. It is now generally accepted, however, that institutions are important in shaping economic development both across space and over time. The policy and intellectual challenges lie in defining and putting in place appropriate institutional structures that meet the objectives of regions and allow them to move

towards these given prevailing constraints not only in terms of the resources available within the region, but also in terms of those imposed by the legitimate desires and actions of other regions. In the 'perfect' world of the neoclassical economist the 'invisible hand' of Adam Smith's market would provide all the stimulation and coordination that is required, and the 'survival of the fittest' à la Herbert Spencer combined with Schumpeterian destructive competition would ensure optimal technical progress. In reality, in a world of imperfect property rights allocation and transaction costs, the invisible hand often requires augmentation by the far from perfect hand of government, broadly defined.

The role of institutions in this sense has always been there – even Stone Age tribes had their leaders and rules – but these have become refined as society has evolved forward, technology has changed, and 'political' structures have evolved. The institutions needed in the modern, knowledge-based society at all levels of spatial interest are not those found in the industrial age, and are unlikely to be those needed in the future. Institutions evolve, both formally and informally, and will continue to influence the rate at which different regions develop and the ways in which this development occurs.

Notes

1. Oliver Williamson (2000) offers a good survey of this literature, and although it is published in an economic journal it takes a much broarder, transdisciplinary perspective. Matthews (1986), in his presidential address to the UK's Royal Economic Society, lays down the argument that all economics is essentially institutional in its orientation today.
2. A fuller definition of institutions is provided by Ostrum (1990): 'Institutions . . . define the sets of working rules that state who is allowed to make decisions in some arena, what actions are allowed or constrained, what aggregation rules are allowed . . . what payoffs can be provided . . . the working rules are used, monitored and enforced when individuals make choices".

   Schmoller (1900, quoted in Furubotn and Richter, 1997) views institutions as providing a basis for steering social actions over long periods of time, for example market system, coinage system, freedom of trade, property, and so on.
3. The coverage here must of necessity be limited but Matthews (1986) offers an accessible account of the economic role of institutions in macroeconomic growth.
4. For a rigorous critique of this framework see Leibenstein (1979).
5. The theoretical outcomes obtained from assumptions of a frictionless world are unlikely in the real world. As Stigler (1972; quoted in Furubotn and Richter, 1997) noted: 'the world of zero transaction costs turns out to be as strange as the physical world would be without friction'.
6. Knight (1922) stressed the need for analyzing 'human nature as we know it' and identified 'moral hazard' as an endemic condition with which economic organization must contend (quoted in Furubotn and Richter, 1997).
7. With its assumptions, neoclassical theory can not differentiate between institutions which are really different – for example between a socialist economy and a private-ownership economy, or say between a money-using economy.
8. The concepts of bounded rationality and satisficing are most closely associated with the work of Herbert Simon; for an outline of his ideas see Simon (1955).
9. In this context, Adam Smith's pin factory captures the idea of intra-firm transactions whenever a pin changes hands in a factory. In the case of the broader market, transactions occur, according to Smith's notion, the division of labor is limited by the extent of the market.
10. Some economic historians who apply NIE concepts such as transaction costs, property rights and contractual relationships to historic economic experience attempt to make institutions endogenous variables in their economic models (North, 1990, 1994).
11. Some analysts have suggested that the rise of a related economic organization of the factory system in the same century was a response by owners of capital to restrain opportunism and realize savings in the process of bargaining with labor (Williamson, 1985; Marglin, 1974).
12. The state realizes a monopoly of organized violence that allows it to institute and run the various organizations necessary to monitor, create and enforce contracts, property rights transactions involved in economic transactions (Levi, 1988).

13. The institutional challenge is not only one of defining the laws and regulations that pertain, but at the regional level of defining appropriate jurisdictions. This, for example, posed problems when the US was formed in terms of federal and state authority, and more recently led to the notion, still being refined in the European Court of Justice, regarding 'subsidiarity' within the European Union.
14. The institutional issue here is to demonstrate that state intervention can improve on the status quo. In some cases, often because of imperfect information, state involvement may worsen the situation – a 'policy intervention failure'.
15. Sabel (1997), Florida (2002) and others suggest that structures other than the market or the firm become necessary in these cases.
16. According to Dasgupta (1988), trust involves correct expectations about the actions of other actors who affect one's action decision, when such a decision has to be made before one is able to monitor the actions of those others. The inability to monitor others' actions is crucial here.
17. The resulting learning, knowledge accumulation and economic evolution are helped in such regions by two kinds of proximities: geographical, leading to reduction of production and transaction costs; and socio-cultural – shared behavioral, cognitive and moral codes (Camagni, 1994). This combination of geographic and socio-cultural proximities leads to a greater intensity of interactions, limited opportunistic behavior, greater division of labor, and cooperation among the enterprises. The operation of such networks and the constituent learning processes are elaborated in Florida (2002).
18. These structural conditions include: reversability of transactions, no scale economies, complete contingent markets, and absence of uncertainty
19. We have not analyzed this institution in this chapter, and present it here to elucidate the relationships among the different institutions.

References

Alter, C. and J. Hage (1993), *Organizations Working Together*, Newbury Park, CA: Sage Library of Social Research 191, Sage Publications Inc.

Baumol, W.J., R.E. Litan and C.J. Schramm (2007), *Good Capitalism, Bad Capitalism and the Economics of Growth and Prosperity,* New Haven, CT: Yale University Press.

Boyer, R. (1997), 'The variety and unequal performance of existing markets: farewell to Doctor Pangloss', in J.R. Hollingsworth and R. Boyer (eds), *Contemporary Capitalism*, Cambridge: Cambridge University Press, pp. 55–93.

Camagni, R. (1994), 'Du milieu innovateur aux réseaux globaux', *Le Courrier du CNRS*, **81**, 36–7.

Chandler, A.D., Jr. (1977), *The Visible Hand: The Managerial Revolution in American Business*, Cambridge, MA: Harvard University Press.

Coase, R.H. (1937), 'The nature of the firm', *Economica*, **4**, 386–405.

Coase, R.H. (1960), 'The problem of social cost', *Journal of Law and Economics*, **2**, 1–40.

Coase, R.H. (1988), *The Firm, the Market, and the Law*, Chicago, IL: University of Chicago Press.

Coleman, J.S. (1990), *Foundations of Social Theory*, Cambridge, MA: Harvard University Press.

Commons, J.R. (1934), *Institutional Economics*, Madison, WI: University of Wisconsin Press.

Dasgupta, P. (1988), 'Trust as a Commodity', in D. Gambetta (ed.), *Trust: Making and Breaking Cooperative Relationships*, Oxford: Basil Blackwell, pp. 49–72.

Doloreux, D. and S. Parto (2005), 'Regional innovation systems: current discourse and unresolved issues', *Technology in Society*, **27**, 133–53.

Enright, M. (1998), 'Regional cluster and firm strategy', in A.D. Chandler, Ö. Sölvell and P. Hagström (eds), *The Dynamic Firm: The Role of Technology Strategy, Organization and Regions*, Oxford: Oxford University Press, pp. 315–42.

Florida, R. (2002), *Rise of the Creative Class: How it is Transforming Work, Leisure, Community and Everyday Life*, New York: Basic Books.

Furubotn, E.G. and R. Richter (1997), *Institutions and Economic Theory*, Ann Arbor, MI: University of Michigan Press.

Granovetter, M. (1985), 'Economic action and social structure: the problem of embeddedness', *American Journal of Sociology*, **91**, 481–510.

Hayek, F.A. (1973), *Law, Legislation, and Liberty*, Chicago, IL: University of Chicago Press.

Hollingsworth, J.R. and R. Boyer (1997), *Contemporary Capitalism*, Cambridge: Cambridge University Press.

Knight, F.H. (1922), 'Ethics and the economic interpretation', *Quarterly Journal of Economics*, **36**, 454–81.

Leibenstein, H. (1979), 'A branch of economics is missing: micro-micro theory', *Journal of Economic Literature*, **17** (3), 477–502.

Levi, M. (1988), *Of Rule and Revenue*, Berkeley, CA: University of California Press.

Marglin, S. (1974), 'What do bosses do? The origins and functions of hierarchy in capitalistic production', *Review of Radical Political Economics*, **1** (6), 60–112.

Marshall, A. (1920), *Principles of Economics*, 8th edn, London: Macmillan.
Matthews, R.C.O. (1986), 'The economics of institutions and the sources of growth', *Economic Journal*, **96** (384), 903–1018.
Mill, J.S. (1857), *Principles of Economics*, 4th edn, London: Parker.
Nelson, R.R. (1959), 'The Simple Economics of Basic Scientific Research', *Journal of Political Economy*, **67** (3), 297–306.
Nelson, R.R. (ed.) (1982), *Government and Technical Progress; A Cross Industry Analysis*, New York: Pergamon Press.
North, D.C. (1990), *Institutions, Institutional Change and Economic Performance*, Cambridge, MA: Cambridge University Press.
North, D.C. (1994), 'Economic performance through time', *American Economic Review*, **84**, 359–68.
Ostrom, E. (1986), 'An agenda for the study of institutions', *Public Choice*, **48**, 3–25.
Ostrom, E. (1990), *Governing the Commons: the Evolution of Institutions for Collective Action*, Cambridge: Cambridge University Press.
Polanyi, K. (1957), *The Great Transformation: The Political and Economic Origins of our Time*, Boston, MA: Beacon Press; originally published 1944.
Polanyi, M. (1983), *The Tacit Dimension*, New York: Anchor Books; originally published 1967.
Posner, R.A. (1977), *Economic Analysis of Law*, Boston, MA: Little, Brown.
Romer, P. (1986), 'Increasing returns and long-run growth', *Journal of Political Economy*, **94**, 1002–37.
Romer, P. (1990), 'Endogenous technological changes', *Journal of Political Economy*, **98** (5), part 2, S71–102.
Romer, P. (1994), 'The origins of endogenous growth', *Journal of Economic Perspectives*, **8** (1), 3–22.
Sabel, C.E. (1997), 'Constitutional orders: trust building and response to change', in J.R. Hollingsworth and R. Boyer (eds), *Contemporary Capitalism*, Cambridge: Cambridge University Press, pp. 154–87.
Simon, H. (1955), 'A behavioral model of rational choice', *Quarterly Journal of Economics*, **69** (1), 99–118.
Sjostrand, S.-E. (1995), 'Towards a theory of institutional change', in J. Groenewegan, C. Pitelis and S.-E. Sjostrand (eds), *On Economic Institutions*, Aldershot, UK and Brookfield, US: Edward Elgar, pp. 19–44.
Stigler, G.J. (1972), 'The law and economics of public policy', *Journal of Legal Studies*, **1**, 1–12.
Stiglitz, J.E. (1985), 'Information and economic analysis: a perspective', *Economic Journal Supplement*, **95**, 21–41.
Williamson, O.E. (1975), *Markets and Hierarchies*, New York: The Free Press.
Williamson, O.E. (1985), *The Economic Institutions of Capitalism*, New York: Free Press.
Williamson, O.E. (2000), 'The new institutional economics: taking stock', *Journal of Economic Literature*, **38**, 595–613.

23 Regional policy: rationale, foundations and measurement of its effects

Jouke van Dijk, Henk Folmer and Jan Oosterhaven

23.1 Introduction

Since the 1930s generations of policy-makers have developed and implemented regional policies for both economic (efficiency) and social (equity) reasons. As regards efficiency, regional disparities in, for instance, unemployment and per capita income often have negative effects on the efficient operation of the national and regional economy. Armstrong and Taylor (2000) give several arguments. The whole nation is better off when unemployed in regions with high unemployment become employed, if the possible loss of jobs in other regions is smaller. A redistribution of jobs may lower the number of hard to fill vacancies in regions with low levels of unemployment. Other benefits are an increase in gross domestic product (GDP) at the national and regional level, and lower costs of social security. A more equal spread of economic activities may also reduce the negative cost of congestion, such as traffic jams and environmental damage in the more densely populated regions of a country. Finally, smaller regional differences in unemployment may also reduce inflationary pressure. As regards equity, reducing interregional disparities may contribute to the general objective of reducing all kinds of unwanted inequality between individuals. In this respect, two classical dilemmas (Stilwell, 1972) are still relevant.

First, there is the dilemma of 'place prosperity versus people prosperity'. At first instance, a direct targeting of individual inequities by means of, for instance, income support seems the preferred strategy. Such social security programs may also contribute to interregional equity, as their recipients tend to be over-represented in the lagging regions (see Stoffelsma and Oosterhaven, 1991; Huffman and Kelkenny, 2007). However, 'place prosperity' may still be needed as an independent goal alongside 'people prosperity', as pursuing only the latter may have unwanted indirect effects. The most important of these are the negative effects of cumulative outmigration of (re)schooled and entrepreneurial individuals, which thus aggravate the situation of the less-schooled, low-income part of the population staying behind. Policy measures that enhance the place characteristics of the region, for instance by means of building new infrastructure, will be mainly beneficial for the individuals that stay in the region. On the other hand, using regional policy for social purposes assumes that this helps the poor individuals in poor regions and, thus, works in the same direction as non-spatial social security policies. However, Dupont (2007) argues that there is little theoretical and empirical evidence for this assumption, and shows that policies designed to reduce interregional disparities may well increase individual inequality. Duranton and Monastiriotis (2002), in fact, show this has been the case in the United Kingdom for the period 1982–1997. Thus, this first classical dilemma is still unresolved.

The second dilemma regards the issue of 'interregional equity versus national efficiency'. Richardson (1979) reviews the neoclassical foundation of the efficiency–equity

trade-off hypothesis and lists cases of efficiency–equity compatibility, such as generative (now labeled endogenous) growth policies and growth pole policies, which may promote interregional equity in an efficient manner. The Williamson (1965) curve, with interregional equity occurring at low and at high levels of economic development, represents another case of efficiency–equity compatibility. Note that this case should not be interpreted as a predecessor of the Krugman (1991) model with spreading equilibriums at high and at low levels of transport cost, as the agglomeration/spreading of workers and firms is not the same as the divergence/convergence of real wages, per capita incomes or welfare levels (see, for a more detailed explanation, Oosterhaven, 1997).

In order to provide a foundation for the discussion of these and other dilemmas and to understand the logic behind regional policy measures, we will discuss several theories that underpin the choice between different regional policy strategies in section 23.2. The subsequent choice between different types of regional policy instruments and an overview of these instruments will be discussed in section 23.3. Note that there does not exist a one-to-one correspondence between theories and instruments, because theories partly overlap and instruments can sometimes be based on more than one theory. As the choice of policy instruments will be based on estimates of their effects, section 23.4 discusses the measurement of the effects of aggregate regional policy and of individual regional policy instruments.

23.2 Theories of regional development

Regional policy is usually founded on theories of regional development. Armstrong (2002) states that there are at least seven related theories of regional growth that play a role in formulating regional policy. Although Armstrong does not pay attention to some recent developments, like the work of Florida (2002) about the 'creative class' and the emerging literature on evolutionary economic geography (Boschma and Lambooy, 1999), we will use his classification as a framework for a brief overview of the relevant theories of regional economic growth.

Neoclassical growth theory

In the interregional versions of this theory, output growth is determined by the growth and mobility of production factors and technology (see Capello, 2007). It predicts that in the long run regions converge and regional per capita GDP disparities will disappear. Convergence occurs because lead regions accumulate capital faster till they run into a situation of diminishing returns that makes investment in lagging regions more attractive and productive. This process is reinforced by four other convergence mechanisms: interregional trade, labor migration, capital mobility and technology transfer. Typical policy instruments based on this theory are the stimulation of labor mobility, free trade and technology transfers.

Endogenous growth theory

An important shortcoming of neoclassical growth theory is that technological progress is assumed to be exogenous. The main feature of endogenous growth theory, as developed by amongst others Romer (1986, 1990), is that technological progress is explicitly modeled, and is itself determined by the growth process. Depending on the way in which technological change is made endogenous (key aspects are human capital, scale effects, spillovers

from investment in physical capital and R&D, and the provision of public services) the outcome can be convergence, but may also lead to cumulative polarized growth. Recent empirical papers analyze the links between growth, geography, agglomeration and learning spillovers (see, for instance, Autant-Bernard et al., 2007). They show evidence of localized knowledge spillovers. Jaffe et al. (1993), for example, show that new patents generally cite previous patents from the same geographical area. Ciccone and Hall (1996) find a positive link between density and productivity of firms in the USA, whereas Broersma and Van Dijk (2005, 2008) find evidence for the Netherlands that high density can also be a disadvantage when congestion and shortages of local production factors, such as land, hamper productivity growth. Typical policy instruments are increasing the level of education of the labor force and the stimulation of start-ups, spin-offs and knowledge diffusion.

Post-Fordism and 'radical' theories

Post-Fordism views history as a sequence of periods of conflicts and consensus between the working and the capital-owning classes. In the post-Fordism model of production, technological change offers firms the opportunity to trade their products globally and enjoy economies of scale, but also requires flexible production methods in response to changing consumers' fashions. This can be realized within a geographical concentration of small and medium-sized firms ('new industrial districts'). Regions which are able to develop such new industrial districts will be booming, while those which are not will stay behind. How long this lasts is not a priori clear, because neoclassical convergence processes also continue to exist (Dunford and Smith, 2000; Glaeser and Gottlieb, 2006).

Social capital theory

This theory emphasizes the impacts of social, cultural and political influences on economic growth, although the focus is more on networks and social cohesion. It has come to the foreground in regional science since Putnam (1993) used it to explain the large differences in income levels between northern and southern Italy. Social capital as such can be added as an extra production factor in the framework of the neoclassical growth theory. In the regional policy debate, however, social capital theory is mainly used to motivate policy measures that develop social capital in lagging regions as a goal itself, whereas the ultimate goal is of course to stimulate economic growth. Durlauf (2006) argues that although there is a strong interest in social capital in economics, the concept itself has proven to be too vague to permit analysis with clarity and precision that matches the standards of the field. This criticism has been developed in a spatial context by among others Florida (2002) and Westlund (2006).

New economic geography models (NEG)

NEG models are based on the work of Krugman (1991) and are essentially cumulative causation models (see Ottaviano and Puga, 1998; Neary, 2001, for overviews). Once a region has got a head start, it attracts new firms and labor because it is able to exploit economies of scale and variety. The agglomeration process can also be driven by productivity effects from close input–output linkages (Venables, 1996). The cumulative causation process may lead to increasing regional disparities, but when transport costs fall sufficiently, convergence is also a possible long-run outcome. Adding congestion costs produces more cases of long run dispersed equilibriums (Brakman et al., 2001). This

theory is typically pessimistic about the effects of policy and it gives no recommendations for policy measures.

Evolutionary economic geography (EEG)
In EEG, agglomeration advantages are also important, but it focuses much more on the role of entrepreneurship and innovation in the Schumpeterian sense in relation to the cohesion in networks and clusters (Boschma and Kloosterman, 2005). EEG differs from NEG and neoclassical theory in that it assumes bounded rationality. It focuses on the explanation of processes of change in which a region is seen as a 'complex adaptive system' wherein the generation and use of knowledge is a crucial factor. Technology is seen as a combination of knowledge and competences. Knowledge is subdivided into 'information' (data), 'coded knowledge' (books, websites, patents, and so on) and 'tacit knowledge' (embedded in persons). Information and coded knowledge become more easily available and distance becomes much less important due to technological progress. The accumulation and use of tacit knowledge is still, or even more so, influenced by geographical proximity.

Demand-driven export competition models
The essential mechanisms in these models are that some regions are more competitive in export markets than others. Increasing competitiveness is mostly based on Verdoorn's Law, where productivity growth is a function of growth of total output. More recently Porter (1990) has added that competitive strength is likely to occur in regions where four mutually reinforcing elements are present: good factor conditions like skilled labor; a strong set of related supporting industries; a competitive milieu for firms within the region; and a strong and critical local demand. Both models predict a cumulative causation process that leads to divergence of regions, as some regions are more successful in creating clusters of exporting firms than others.

Innovative milieus and 'learning' regions
Several of the previous theories take the emergence of a geographical cluster of high-tech firms ('innovative milieux') as a factor that causes divergence between regions. The innovative milieu theory presents the underlying mechanisms. In such milieux firms develop and retain key competencies necessary for rapid growth and success (Lawson, 1999). Of special importance is a pool of specialized labor that shares and combines knowledge within a complex system, and forms and maintains effective social relations in organizations. This implies that such regions become 'learning regions', which are attractive for dynamic people and firms, and will therefore show higher growth rates than other regions. This argument typically fits with the ideas of Florida (2002) about the importance of the 'creative class' for regional development. Urban regions that are attractive to dynamic people and firms will have dynamic workers ('the creative class') and entrepreneurs and will, therefore, produce higher growth rates than other regions (Audretsch et al., 2006). Saxenian (2006) adds that the globalization of production systems and processes of outsourcing requires an emerging group of entrepreneurial knowledge workers ('new Argonauts') that are internationally mobile. Regions with the appropriate production structure and an open innovation system that are attractive to these 'new Argonauts' tend to show higher growth rates (Atzema and Boelens, 2006)

From the above overview it is clear that regional disparities in income levels and growth rates may be explained by a broad variety of relevant factors. Some theories predict that regional disparities will converge over time, while others predict divergence. Most theories, however, allow for different outcomes under different conditions. Attempts to explain empirically why some regions succeed (and others do not) have by and large identified similar factors, although the terminology may vary from one study to another. In a recent attempt to explain differences in economic performance between European regions, Cuadrado-Roura (2001) identifies seven attributes that correlate positively with superior performance. In our opinion they adequately summarize the current state of the empirical outcomes. The factors are:

1. City system: the presence in a region of a group of medium-sized cities (population: 40 000 to 150 000) in combination with a large city.
2. Human resources: supply of labor with medium to high educational levels, preferably with moderate wages.
3. Accessibility: proximity to major markets and large urban centers in a physical sense, but also in terms of access and receptivity to new ideas.
4. Producer services: a varied set of firms specialized in consulting, advertising, finance, and so on.
5. Institutional infrastructure: a supporting local government with well-developed development strategies and leadership from the region itself.
6. Image: a positive social climate (particularly, few labor conflicts) and a local environment conducive to cooperation among institutions and organizations.
7. Industrial size mix: many small and medium-sized firms easily leading to knowledge spillovers, as opposed to dominance of a region by a few large firms.

Of course, the characteristics of regions that correlate with relative, and eventually absolute, population and employment decline are also important from a policy point of view. Based on a study of Canadian regions Polèse and Shearmur (2006) formulate the following preconditions for regional decline.

National attributes:

1. A geographically large nation with a periphery, that is, inhabited spaces located beyond a one-hour drive from a major urban center.
2. A nation in the last stages of demographic transition: natural population increase is either close to zero or negative.

Regional attributes:

1. Located in the periphery of the national or continental space economy.
2. Not located along a major transport axis or trade route.
3. No urban area of over 100 000 and/or less than three urban areas with over 40 000 within a 100 km range of each other (thresholds can vary with context).
4. An economic base in resource exploitation and/or primary processing.
5. A resource base whose limits of (profitable) exploitation have been reached.

6. Presence of Weberian weight-losing[1] industries, capital-intensive, with high labor productivity and high wages.
7. Climatic and geographical conditions that limit year-round tourism.

All of the conditions need not be present for decline to occur. In poorly located, sparsely populated regions, attribute 6 for instance is not a necessary condition. Moreover, attributes may reinforce each other, but a positive attribute may also compensate negative impacts of others.

What are the implications of both lists? One could argue that policy measures for regions destined to decline are a waste of resources given the very low chances of success. For regions that meet the criteria for success, regional policy measures do not seem necessary because these regions can realize growth without help. However, most regions will have characteristics in–between these two extremes, and regional policy measures could be aimed at removing less favorable characteristics and stimulate the creation of success factors.

23.3 Instruments of regional policy

Let us now turn in more detail to the implications of the above for the design of regional policy measures. From a people's prosperity perspective, it seems most evident to help immobile unemployed in lagging regions to move to more prosperous ones. The rationale is that if individuals were perfectly mobile the unemployment problem would not persist. Apparently there are major obstacles to mobility. In this respect there are interesting differences between Europe and the US. The classic studies by Blanchard and Katz (1992) for the US and by Decressin and Fatás (1995) for Europe show that there are major differences in the adjustment process in the regional labor markets. In the US about two-thirds of the adjustment takes place via migration, whereas in Europe changes in regional participation rates account for more than two-thirds of the adjustment. This result has been confirmed in many studies and is remarkably stable (see Broersma and Van Dijk, 2004). The high mobility of workers may explain why regional policies and regional inequality are not considered public priorities in the US (Dupont, 2007). In the US, regional policy mainly takes the form of stimulating job creation and entrepreneurship at the city level, and of reducing racial and class segregation (Jonas and Ward, 2002).

In Europe, in contrast to the US, migration stimulation policy measures have been used in the 1960s and 1970s. Van Dijk (1986) shows for the Netherlands that the number of workers assisted was very limited. Moreover, Van Dijk et al. (1989) show that the correlation between migration rates and re-employment rates is much weaker in the US than in the Netherlands. The main explanation is that unemployed in the Netherlands migrate after they have found a job (contracted migration), whereas in the US they do so before (speculative migration). In a follow-up study, Van Dijk et al. (2000) show that the benefits of migration in the US are especially high in terms of wage increases for highly skilled workers, whereas such benefits are much lower in most continental European countries due to their centralized national wage-setting. Layard (2006) argues that stimulation of migration does not seem to be an adequate instrument to help people with poor chances on the labor market, as their willingness (and ability) to move to other regions is low. Moreover, he argues that although migration may lead to higher income, the negative effects of loss of family stability and higher crime rates tend to dominate the income gain.

Instead of moving 'workers to work', moving 'work to workers' is a different and much more used regional policy strategy in Europe. In this respect there are two approaches:

1. moving jobs from regions with tight labor markets to regions with high unemployment rates (exogenous or redistributive growth);
2. stimulating the creation of new jobs in regions with high unemployment rates (endogenous or generative growth).

Moving jobs can take several forms. Especially in the UK in the 1960s and 1970s, firm migration was seen as a means to transfer work and prosperity to lagging regions, while at the same time reducing congestion, such as labor shortage and lack of space, in the core regions. In the 1980s and 1990s, however, policy incentives were much more focused on the stimulation of endogenous growth, while policy-makers lost interest in stimulating firm migration (Pellenbarg et al., 2002). Another way of moving jobs from leading to lagging regions is the relocation of public sector jobs. Oosterhaven (1981) has shown that this has been a very effective regional policy measure for the Netherlands, and recently Marshall et al. (2005) point out the positive effects of public sector dispersal for Britain.

By far the most important strand of regional policy measures nowadays aims at stimulating job creation in lagging regions. Over time and space, a host of different types of measures have been taken, such as supporting starting firms, providing export and innovation subsidies for small local firms, creating social networks among firms, stimulating spin-offs of universities and technological institutes, and providing office space with common facilities. Some measures are in operation for only a few years, but others are applied for longer periods. The differences in measures reflect the continuously changing ideas about how regional development can best be achieved. Oosterhaven (1996) presents the following typology of policy measures to stimulate job creation:

1. direct versus indirect measures;
2. restrictive (stick) versus stimulating (carrot) measures;
3. subsidies on capital versus subsidies on labor;
4. single measures versus integrated packages.

Direct versus indirect measures

To stimulate firm establishments a government can opt for indirect measures, such as creating a better infrastructure (roads, railroads, industrial sites, and so on), stimulating research and development (R&D) and innovations, improving educational facilities and providing an attractive environment (housing, facilities for sport and recreation, cultural facilities, and so on). Such improvements may tempt entrepreneurs to locate in an area they used to ignore as a location alternative. Alternatively, government can choose a more direct way by offering financial compensation such as tax deductions (fiscal zones), soft loans, investment premiums, low land prices, favorable energy contracts, and so on. These measures are more direct than investments in infrastructure, and so on, as they provide a direct financial contribution to an individual firm, but they are also indirect because the effect ultimately depends on the entrepreneur who is free to decide. Examples of strictly direct measures are the relocation of government institutions, which is under complete control of the government, and financial assistance of specific firms who are in financial

trouble. Substantial financial participation in firms (over 50 percent) by state-owned regional development companies may also be seen as a rather direct policy measure.

The stick versus the carrot

Relocation of economic activity is influenced by push- and pull-forces. Attempts to influence relocations are mostly aimed at the pull side by making the target regions as attractive as possible. Investment premiums are a good example of a carrot to attract new firms or to stimulate growth of existing firms. But it is also possible to act on the push side by stimulating entrepreneurs to outmigrate from congested areas with tight labor markets. This can be done by direct push-measures, such as permits for investing in new establishments or expanding existing ones, or by more indirect push-measures, such as congestion taxes.

Subsidies on capital or labor

When the carrot is chosen, the next question is: which production factor(s) should be subsidized: labor, capital, energy, intermediate inputs, services, knowledge? The choice will depend on such factors as output, price, substitution possibilities, sector structure, multiplier and long-run effects. This can be illustrated by the choice between capital and labor subsidies (see also Armstrong and Taylor, 2000). Labor subsidies will stimulate the output of labor-intensive industries or the remuneration of labor. Moreover, labor subsidies may lead to substitution of capital by labor. In the case of capital subsidies particularly capital-intensive industries and profits will benefit. Labor may be substituted by capital leading to a decrease in the number of jobs. So, when employment creation is the primary goal, labor subsidies are more effective than capital subsidies. This effect on labor is further strengthened by indirect macro effects, as increased labor demand and higher wages will lead to local expenditure which, *ceteris paribus*, is greater than the effects due to subsidized capital and profits, since the latter impacts are more likely to end up outside the region, especially when it is weak and lagging.

There are two arguments that qualify this conclusion. First, labor-intensive industries may have less long-run growth potential than capital-intensive industries, especially in the case of weak regions in developed countries. Second, it is more effective only to subsidize an idle production factor rather than (also) its stock already in use which is more easily implemented in the case of capital than of labor. Of course, one should avoid subsidizing mobile capital, such as means of transportation, that can easily leave the region after the subsidy is collected. The implementation of a subsidy on new jobs is more complex than that on new capital: what is a new job and for how long should the subsidy last? Comparable, but differently weighted, arguments apply to the choice for technology, R&D, energy and other factor or sales (exports) subsidies.

Single measures versus integrated packages

In its first stages a regional policy program tends to focus on specific policy goals and specific policy measures. Later on (partly due to bureaucratic forces) regional programs usually develop towards complex packages of policy goals and policy measures. The costs of these packages tend to increase strongly while it becomes more and more difficult to evaluate their effects, and especially the effects of single measures included in these packages. Therefore, when funds for regional policy are limited, it is advisable to restrict the

set of measures and make them as selective as possible, aimed at a restricted set of goals or even a single goal (see Oosterhaven, 1996, for a description of this long-term development in regional policy in the Netherlands).

23.4 Measurement of effects

Measurement of policy effects is a crucial step in the development and adaptation of regional policy. In this section we present a brief overview of the main measurement approaches and methods. We start with the conceptual framework which is made up of the following elements:

1. Policy goals and targets: as mentioned above, the most general goal variables of regional policy are equity and efficiency. From these policy goals more concrete and operational policy targets need to be derived, such as the desired level of investment or employment growth.
2. Policy instruments, that is, the variables under the command of the government by which the policy goals and targets are pursued. An instrument consists of a specific set of governmental acts which are internally cohesive and externally distinguishable.
3. Non-policy variables: the policy targets are usually influenced not only by policy instruments but also by non-policy variables that are not under the command of the regulator. The effects of the non-policy variables may be more important than the effects of the policy instruments.
4. Additional impact variables: regional policy may not only affect policy targets but also other variables.
5. Direct and indirect effects: indirect effects on policy targets materialize via intermediate variables.
6. Time lags: some effects may materialize in the short run whereas others materialize over longer time horizons. Moreover, one and the same effect may materialize over several periods.
7. The total effect of a policy instrument consists of all the direct and indirect effects on both the policy and additional impact variables over all time periods.

The measurement approaches can be divided into micro and spatial approaches. In the former the data are of the lowest level of aggregation and refer to such units as firms, households, workers, and so on. In spatial approaches the individual data have been aggregated by regions.

Micro approaches

Controlled experiments This type of micro approach has been used relatively little in impact studies of regional policy so far. There seems to be a growing interest, however, and therefore it is described here. The essence of laboratory experiment lies in observing the effect on a dependent (policy target) variable of the manipulation of an independent variable under controlled conditions. Rather than this extreme form, a more common approach in (regional) economics operates along the following lines. Groups with different histories of the policy instrument are selected, and in these groups differences with respect to the policy target are analyzed. The selection is supplemented by matching as many of

the relevant independent policy and non-policy variables as possible, so as to control the effects of other variables than the policy instrument. In practice, only a limited number of variables can be matched. Therefore, the matching is supplemented by 'control through measurement'. This means that a relevant variable, which cannot be used for matching, is taken into account by gathering information about it from the respondents. Finally, the possible disturbing influences of uncontrolled variables are taken into account by randomly selecting samples from the matched groups (see, among others, List, 2006).

It is clear that the main advantage of controlled experiments lies in the high degree to which the causal relations between policy instrument and policy target are isolated from the disturbing influences of other variables. There are, however, various difficulties inherent in the use of controlled experiments in policy research. First, the problem of matching may be difficult or impossible. In regional impact studies the problem of matching is worsened by the fact that both the experimental and control groups have to be located in the same region or at least in similar regions, which may seriously limit the number of members in both groups. Secondly, the results are limited by the experiences of the groups involved in the experiment. So the specific results may be difficult to generalize to other situations. Finally, the experimental setting may have disturbing effects on the participants, which may lead to unreliable outcomes. In spite of these difficulties, controlled experiments have been applied in various situations. For instance, Smith (1979) studied the provision of public goods while Hall (1975) analyzed the effects of income taxes on the supply of labor.

In general economics controlled experiments have become a standard research tool and are applied in a wide variety of fields including market behavior, decision-making, bargaining, social preferences, learning, free-riding, the provision of public goods and environmental policy. See *inter alia* Kagel and Roth (1995). Since many of the topics mentioned above are also relevant in regional science, it would be worthwhile investigating the applicability of controlled experiments in this field.

Quasi-experimental and non-experimental research Quasi-experimental research consists of surveys among agents who are expected to have been affected by policy. They may provide detailed information on the various factors influencing decision-making and especially on the relative weights of the policy instruments. The information obtained via surveys may relate to direct and indirect effects of policy. As an example of the latter, consider the case that investment subsidies have led to the establishment of an important industrial enterprise in a given region. In order to assess indirect effects for other firms, a survey could be held among these firms with respect to the importance for their operations of the newly located enterprise. From the above it follows that well-designed surveys may give detailed information about the decision processes of the respondents; in particular, of their perceptions of the importance of relevant factors. Furthermore, surveys may provide information to make comparisons between different situations, for example before and after the move of a firm. Finally, information about such matters as time lags between decisions with regard to, and realizations of, for example, investments may be obtained.

The survey approach as a measurement method may suffer from the drawbacks of surveys in general. These can be grouped under the headings: misinterpretation of questions, strategic responses and measurement errors. An example of the first problem

is the *ex post* rationalization of the proper factors underlying the decision made. An example of the second problem is the under-reporting of the effect of investment subsidies to promote higher subsidies in the future. Part of these problems may be mitigated by incorporating questions which only indirectly relate to the policy variables. Another limitation is that the information obtained is affected by the perceptions and expectations of the interviewees. Moreover, it is often difficult to obtain reliable data on the past because of the 'loss of memory' of organizations as a consequence of new management, destruction of information, and so on. The survey method has been applied to a variety of problems. For instance, Baumont (1979) investigated the impacts of migration incentives; Calame (1980) studied effects of wage subsidy programs; Marquand (1980) and Krist (1980) investigated the impacts of investment subsidies on the location decisions of firms; whereas Pellenbarg et al. (2002) and Meester and Pellenbarg (2006) used the approach to analyze firm migration and the spatial preference map of Dutch entrepreneurs, respectively.

The second kind of micro approaches consists of non-experimental research. Whereas the researcher has control over the influences of the causal variables on the policy target in experimental research, and over data collection in quasi-experimental research, no control is exerted by them in non-experimental research. It is restricted to the observation of the policy target in different situations, such as before and after a policy intervention or in different regions. In non-experimental research no attempts are made to separate effects of policy instruments from effects of non-policy variables. Therefore, its use is restricted to situations where the latter types of effects do not exist or can be taken into account in other ways. This is *inter alia* the case in situations where direct effects of instruments of the control type are under study. For instance, the direct employment effects of the construction of an infrastructural project can be derived from the construction expenses in a straightforward way.

Regional approaches

The data used in spatial approaches are usually obtained from micro units in surveys conducted by authorities, such as a central office of statistics. The surveys tend to be relatively simple and relate to key issues such as investments, number of persons employed, and so on. In contrast to the surveys discussed above, the information asked for usually does not directly relate to regional economic policy. Therefore, there is less danger of answers which have been biased to influence future policy. It is obvious that when no information on policy is gathered from the micro units it has to be obtained elsewhere, for instance at the ministry responsible for the policy. An important advantage of many surveys organized by public authorities, such as a ministry or a central bureau of statistics, is that they are repeated periodically. This means that information to estimate effects of policy becomes available for much longer periods than in the case of the micro studies, which are usually organized only incidentally. However, there are two shortcomings: (1) in repeated surveys only successful firms are tracked over time which may result in survival bias; and (2) not all surveys are representative at the regional level. The former type of problem can be handled by means of careful comparison of drop-outs and survivors, whereas the second requires action at the stage of the sample design.

Using the present kind of data, the following types of approaches have been used to measure effects of regional policy.

Approaches with explanatory variables of the policy type only These approaches compare the scores of impact variables in 'policy-on' situations with their scores in 'policy-off' situations, and ascribe the differences solely to policy. Thus it has to be assumed that the possible effects of non-policy variables may be neglected. This approach can only be used in situations with prior knowledge about the absence of the effects of non-policy variables or in situations when non-policy effects on the impact variables can be completely isolated from effects of policy, as in the case of the relocation of government offices. In practice these conditions are seldom fulfilled.

Single-equation models with explanatory variables of the non-policy type only This type of approach is based on comparisons of the actual policy-on situation with the hypothetical policy-off situation, where the latter is extrapolated on the basis of non-policy variables only. The gap between the two situations is defined as the effect of policy.

Several variants of this class of models can be distinguished. The simplest variant is the extrapolation on the basis of a univariate time series of the impact variable for the policy-off situation in a single region. It rests on the assumption that the autonomous development of the impact variable in both the policy-on and the policy-off period is the same. This assumption may be violated, especially when a development from a short policy-off period is extrapolated over a long policy-on period. The extrapolation may also be made by means of such methods as seasonal autoregressive integrated moving average approaches (Box and Jenkins, 1976). It may also be based on more simple methods, such as relating the development of the impact variable in a policy region to the development of the impact variable in non-policy regions, or to the development of related variables in the same region which have not been affected by policy. If there is evidence of adequate extrapolation of the impact variable, the method is a simple and easy device to estimate policy effects. It has been used by Begg et al. (1976) to measure effects on investments.

Single-equation models with both policy and non-policy variables In the present context, two kinds of situations will be considered. In the first, information on important non-policy variables is missing but is taken indirectly into account. In the second, information on all relevant variables is available. The method to be used in the first situation will be called 'two-stage time-series analysis'. In order to apply two-stage time-series analysis, a univariate time series of the impact variable for the pre-intervention period and a multivariate time series of the impact variable and the various policy instruments for the policy-on period must be available. The first step is to model the pre-intervention series. Of considerable applicability is the class of multiplicative seasonal autoregressive integrated moving average models (that is, SARIMA models; see Box and Jenkins, 1976). The second step consists of removing the effects of the non-policy variables, estimated on the basis of the pre-intervention series, from the second series. This removal is successful if the relationships between the impact variable and the non-policy variables in the intervention period are the same as in the pre-intervention period. Under the conditions of independence of policy instruments of non-policy variables and an additive model structure, the effects of policy on the impact variable can be estimated by standard techniques from the transformed multivariate time series. For an application of the present measurement approach see Folmer (1986).

When information on both policy and non-policy variables is available, standard approaches with both types of explanatory variables explicitly included (that is, multivariate time-series analysis to a single region, multi-regional or interregional cross-section analysis or a spatio-temporal analysis) can be used to estimate effects of the policy instruments. Examples can be found in, among others, Ashcroft and Taylor (1977) and Heckman (1997). Observe that with single-equation methods only direct effects of instruments on an impact variable can be estimated. In addition, single-equation methods as such do not allow of the estimation of the effects of an instrument on several impact variables. For both purposes, either several single-equation models are required or simultaneous equation methods have to be used. Of the latter we will discuss input–output models and general simultaneous equation models.

Input–output models Input–output models represent sectoral and regional disaggregations of the well-known macroeconomic income–expenditure model, with $Y = C + I + G + X - M$. They especially record the intermediate transactions between industries and industry sales of final goods and services to households, government and exports. Input–output models therefore are typically suited to calculate the indirect effects – for example on value-added, employment or energy use – of exogenous final demand changes. The core impact equation of a typical (type II) interregional input–output model reads as follows (see Oosterhaven, 1981):

$$\Delta v = \mathbf{c'} (\mathbf{I} - \mathbf{A} - \mathbf{Q})^{-1} \Delta \mathbf{f} \tag{23.1}$$

where:
Δv = change in the impact variable, for R regions and J industries;
$\mathbf{c'}$ = RJ row with impact variable per unit of output;
$\mathbf{A}$ = RJ x RJ matrix of intermediate input coefficients;
$\mathbf{Q}$ = RJ x RJ matrix of households' expenditure per unit of output;
$\Delta \mathbf{f}$ = RJ column with changes in exogenous final demand.

Model (23.1) can be generalized in various ways. Usually it is embedded in either a demo-economic model framework with time lags (see Batey, 1985; Oosterhaven and Folmer, 1985) or it is embedded in a regional econometric framework (see Treyz, 1993). From (23.1) it is clear that only policy interventions that can be specified in terms of changes in exogenous final demand or changes in the model's coefficients can be handled by means of input–output models, and then only when the economy does not suffer from supply bottlenecks. When policy interventions primarily work through prices and the supply–side of the economy is restricted, different models need to be used.

The use of the input–output models is limited by the scarcity of data, especially with respect to interregional linkages. For the same reason, the relations in input–output models usually are not quantified by means of conventional econometric methods. Examples of the input–output measurement approach can be found in Moore and Rhodes (1976), where the impacts of labor subsidies are investigated; in Oosterhaven (1981), where the effects of the relocation of governmental offices and a land reclamation project are analyzed; and in Oosterhaven et al. (2001), where Dutch spatial mainport policy is evaluated.

General simultaneous equations models The structural form of the conventional general simultaneous equations measurement model reads as follows:

$$A_0 y_t = \sum_{i=1}^{p} A_i y_{t-i} + \sum_{j=0}^{q} B_j x_{t-j} + \sum_{k=o}^{m} C_k z_{t-k} + \varepsilon_t \tag{23.2}$$

where:
y_t = g-vector with current endogenous variables at time *t*;
y_{t-i} = g-vector with lagged endogenous variables in period *t-i*, i lags;
x_{t-j} = m-vector with exogenous non-policy variables in period *t-j*, j lags;
z_{t-k} = n-vector with exogenous policy instruments in period *t-k*, k lags;
ε_t = g-vector with random disturbances;
A_i = g x g matrix with unknown coefficients corresponding to y_{t-i};
B_j = g x m matrix with unknown coefficients corresponding to x_{t-j};
C_k = g x n matrix with unknown coefficients corresponding to z_{t-k}.

As mentioned above, single-equation approaches are not appropriate to decompose policy effects which arise along causal chains of length longer than one, and to estimate effects on several impact variables simultaneously. Both aspects, however, can be handled by means of simultaneous equations models. In order to estimate the direct effects of an instrument of policy, all impact variables should be incorporated into the model as current endogenous variables (that is, should be included in the vector y_t). Each impact variable should be specified as a function of the instruments of policy and of the other relevant explanatory variables. In order to decompose indirect effects of an instrument on a given impact variable, both the ultimate impact variable and each of the intermediate variables in the causal chain between the impact variable and the instrument of policy should be specified as current endogenous variables. Thus, a causal chain is represented by a system of equations where each causal variable is among the explanatory variables of the variable it directly affects. For an example of a general simultaneous equation model see Folmer (1986).

Social cost–benefit analysis Finally, social cost–benefit analysis (CBA) may be used to evaluate regional policy (see Heyma and Oosterhaven, 2005). In doing so, it is important to realize that the welfare of a certain population is the goal variable in any well-done CBA, and not the government-declared goals of the regional policy at hand. Any CBA consists of the following more or less standard stages (Hanley and Spash, 1994; Hanley, 2000).

Stage 1: definition of the project or policy. This includes defining: (1) the policy-driven reallocation of resources being proposed; and (2) the population whose welfare is to be considered.

Stage 2: identification of project impacts. In the case of, for instance, a new railway this stage would include a listing of all resources used in constructing the railway (concrete, steel, labor hours) as well as effects on local unemployment, traffic movements, local property prices, time saving and accidents, wildlife populations, and impacts on the quality of landscape in the area not picked up by changes in property values (for example Elhorst et al., 2004).

Stage 3: selection of economically relevant impacts. A CBA assumes that society is interested in maximizing the weighted sum of utilities across its members. These depend,

amongst others, on consumption of marketed and non-marketed goods (for example clean air). Benefits will either be increases in the quantity or quality of the goods that generate positive utility, or a reduction in the price at which they are supplied, whereas costs relate to opposite effects. CBA is primarily interested in the net total of all effects. Important measurement problems relate to additionality and displacement (SACTRA, 1999). Additionality implies that benefits should be measured net of any effects that would have occurred without the policy (for instance employment effects due to national economic recovery), whereas displacement relates to crowding-out elsewhere (for example employment decline due to the stimulation of investments in the region).

Stage 4: physical quantification of relevant impacts. This involves determining the physical size of the impacts, and identifying when in time they will occur.

Stage 5: monetary valuation of relevant impacts. In order for physical measures of impacts to be comparable, they must be valued in common units like relative prices generated by markets. A CBA should therefore: (1) predict prices for value flows extending into the future; (2) correct market prices where necessary, for example at imperfect markets; and (3) estimate prices for non-market goods, such as lost nature and life.

Stage 6: discounting of cost and benefit flows. Due to the existence of a market interest rate, impatience and risk, future cost and benefits need to be converted into 'present values' to make them comparable. The value of a future cost or benefit (X) occurring in time t is made 'present' with a discount rate d as follows: $PV = X_t(1 + d)^{-t}$

Stage 7: net present value and redistribution test. The main purpose of CBA is to help select projects and policies which are efficient in terms of their use of resources. The net present value (NPV) test simply asks whether the sum of the discounted gains exceeds the sum of discounted losses. If so, the project can be said to represent an efficient change in resource allocation. Secondary inventories of who is actually losing and gaining need to be added to the NPV test in order to enable the political weighting of the efficiency and equity effects for different income groups.

Stage 8: sensitivity analysis. The NPV test tells us about the efficiency of a policy or project, given a set of data and assumptions. If these data and/or assumptions change, then clearly the results of the NPV test will also change. An essential final stage of any CBA is therefore to conduct sensitivity analysis, that is, recalculating the NPV when key values are changed, such as the discount rate, the physical quantities, market and especially non-market prices, and the project life-span.

23.5 Conclusion

In this chapter, we have surveyed the theoretical foundations of regional policy by reviewing eight groups of regional growth theories, and we have discussed the strategic and operational dilemmas in selecting regional policy goals and regional policy instruments. In addition we have given an overview of the micro and spatial approaches that can be applied to assess the efficiency and equity impacts of regional policy, and indicated the relative strength and weaknesses of these approaches. From the overview it is clear that regional disparities in income levels and growth rates can be explained by a broad variety of relevant factors. Some theories predict that regional disparities will converge over time, while others predict divergence, but most theories, however, allow for different outcomes under different conditions. Although the theories provide an analytical and behavioral

framework for the choice of policy goals and policy instruments, it is also a political choice that is dependent on the institutional and socio-economic setting and the norms and values with regard to expected spatial and labor market behavior of individuals and firms in a given society. The discussion about differences in migration behavior in the US and in Europe shows that causal mechanisms may show substantial differences over time and space. The differences in characteristics of regions with very good and hardly any potential for a successful regional policy indicate that the choice of policy goals and policy instruments should take into account these variations. The spatial scale may also be important. It is likely that the choice of appropriate policy goals and effective and efficient policy instruments will be different for a local government in a local setting than at the national or state level or at the level of the European Union or the United States. This even holds for measuring the effects of regional policy, as causal mechanisms are different, data are different and the interference of non-policy variables will be different in space and time. Hence, one size definitely does not fit all.

Note

1. The weight of the final product is less than the weight of the raw material that goes into making the product.

References

Armstrong, H.W. (2002), 'European Union regional policy: reconciling the convergence and evaluation evidence', in J.R. Cuadrado-Roura and M. Parellada (eds), *Regional Convergence in the European Union. Facts, Prospects and Policies*, Berlin: Springer, pp. 231–72.

Armstrong, H.W. and J. Taylor (2000), *Regional Economics and Policy*, Oxford: Blackwell.

Ashcroft, B. and J. Taylor (1977), 'The movement of manufacturing industry and the effect of regional policy', *Oxford Economic Papers*, **29**, 84–101.

Atzema, O.A.L.C. and L. Boelens (2006), 'Connecting Randstad Holland', in A.F. Koekoek, J.M van den Cammen, T.A. Velema and M. Verbeet (eds), *Cities and Globalisation: Exploring New Connections*, NGS-series, nr 339, Utrecht: Royal Dutch Geographical Society, pp. 61–73.

Audretsch, D.B., M.C. Keilbach and E.E. Lehmann (2006), *Entrepreneurship and Economic Growth*, Oxford: Oxford University Press.

Autant-Bernard, C., J. Mairesse and N. Massard (2007), 'Spatial knowledge diffusion through collaborative networks', *Papers in Regional Science*, **87** (3), 341–50.

Batey, P.W.J. (1985), 'Input–output models for regional demographic-economic analysis: some structural comparisons', *Environment and Planning A*, **17**, 77–93.

Baumont, P.B. (1979), 'An examination of assisted labour mobility', in D. MacLennan and J.B. Parr (eds), *Regional Policy: Past Experiences and New Directions*, Oxford: Robertson, pp. 123–42.

Begg, H.M., C.M. Lythe and D.R. MacDonald (1976), 'The impact of regional policy on investments in manufacturing industry: Scotland 1960–71', *Urban Studies*, **13**, 171–9.

Blanchard, O. and L. Katz (1992), 'Regional evolutions', *Brookings Papers on Economic Activity*, **1**, 1–75.

Boschma, R.A. and R.C. Kloosterman (eds) (2005), *Learning from Clusters*, Berlin and Dordrecht: Springer.

Boschma, R.A. and J.G. Lambooy (1999), 'Evolutionary economics and economic geography', *Journal of Evolutionary Economics*, **9**, 411–29.

Box, G.E.P. and and G.M. Jenkins (1976), *Time Series Analysis: Forecasting and Control*, San Francisco, CA: Holden-Day.

Brakman, S., H. Garretsen and C. van Marrewijk (2001), *An Introduction to Geographical Economics*, Cambridge: Cambridge University Press.

Broersma, L. and J. van Dijk (2004), 'Adjustment to labour market shocks in the light of structural reforms', in E. Boneschansker, J. van Dijk, L.G. Jansma and K.H.A. Verhaar (eds), *Cultural Uniqueness and the Regional Economy*, Leeuwarden: Fryske Akademy, pp. 257–75.

Broersma, L. and J. van Dijk (2005), 'Regional differences in labour productivity in the Netherlands', *Journal of Economic and Social Geography*, **96** (3), 328–37.

Broersma, L. and J. van Dijk (2008), 'The effects of congestion and agglomeration on mfp-growth in Dutch regions', *Journal of Economic Geography*, **8** (2), 181–209.

Calame, A. (1980), 'Impacts and costs of wage subsidy programmes: experiences in Great Britain, Sweden and the USA', paper for the International Institute of Management, Berlin.

Capello, R. (2007), *Regional Economics*, London: Routledge.
Ciccone, A. and R. Hall (1996), 'Productivity and the density of economic activity', *American Economic Review*, **87**, 54–70.
Cuadrado-Roura, J.R. (2001), 'Regional convergence in the European Union: from hypothesis to the actual trends', *Annals of Regional Science*, **35**, 333–56.
Decressin, J. and A. Fatás (1995), 'Regional labour market dynamics in Europe', *European Economic Review*, **39**, 1627–55.
Dunford, M. and A. Smith (2000), 'Catching up or falling behind? Economic performance and regional trajectories in the "new" Europe', *Economic Geography*, **76**, 169–95.
Dupont, V. (2007), 'Do geographical agglomeration, growth and equity conflict?', *Papers in Regional Science*, **86** (2), 193–213.
Duranton, G. and V. Monastiriotis (2002), 'Mind the gap: the evolution of regional inequalities in UK 1982–1997', *Journal of Regional Science*, **42**, 219–56.
Durlauf, S.N. (2006), 'Assessing racial profiling', *Economic Journal*, **116** (515), F402–F426.
Elhorst, J.P., J. Oosterhaven and W.E. Romp (2004), 'Integral cost–benefit analysis of Maglev technology under market imperfections', SOM Report 04C22, University of Groningen.
Florida, R. (2002), *The Rise of the Creative Class*, New York: Basic Books.
Folmer, H. (1986), *Regional Economic Policy*, Dordrecht: Kluwer.
Glaeser, E.L. and J.D. Gottlieb (2006), 'Urban resurgence and the consumer city', Harvard Institute of Economic Research, Discussion Paper 2109.
Hall, R.E. (1975), 'Effects of the experimental negative income tax on labor supply', in J.A Peckman and J. Timparrrel (eds), *Work Incentives and Income Guarantees*, Washington, DC: Brookings Institute.
Hanley, N. (2000), 'Cost–benefit analysis', in H. Folmer and H.L. Gabel (eds), *Principles of Environmental and Resource Economics*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, pp. 104–29.
Hanley, N. and C. Spash (1994), *Cost–Benefit Analysis and the Environment*, Aldershot, UK and Brookfield, US: Edward Elgar.
Heckman, J.J. (1997), 'Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations', *Journal of Human Resources*, **32** (3), 441–62.
Heyma, A. and J. Oosterhaven (2005), 'Social cost–benefit analysis and spatial-economic models in the Netherlands', in F. van Oort, M. Thissen and L. van Wissen (eds), *A Survey of Spatial Economic Planning Models in the Netherlands*, Rotterdam: Ruimtelijk Planbureau/NAi-uitgevers, pp. 155–75.
Huffman, S.K. and M. Kelkenny (2007), 'Regional welfare program and labour force participation', *Papers in Regional Science*, **86** (2), 215–39.
Jaffe, A., M. Trajtenberg and R. Henderson (1993), 'Geographic localization of knowledge spillovers as evidenced by patent citations', *Quarterly Journal of Economics*, **108**, 577–98.
Jonas, A.E.G. and K. Ward (2002), 'A world of regionalisms? Towards a US–UK urban and regional policy framework comparison', *Journal of Urban Affairs*, **24** (4), 377–401.
Kagel, J. and A. Roth (eds) (1995), *The Handbook of Experimental Economics*, Princeton, NJ: Princeton University Press.
Krist, H. (1980), 'An appreciation of regional policy evaluation studies: the case of the Federal Republic of Germany', paper of the International Institute of Management, Berlin
Krugman, P. (1991), *Geography and Trade*, Cambridge, MA: MIT Press.
Lawson, C. (1999), 'Towards a competence theory of the region', *Cambridge Journal of Economics*, **23** (2), 151–66.
Layard, R. (2006), 'Happiness and public policy: a challenge to the profession', *Economic Journal*, **116** (March), C24–C33.
List, J.A. (ed.) (2006), *Using Experimental Methods in Environmental and Resource Economics*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar.
Marquand, J. (1980), 'Measuring the effects and costs of regional incentives', Working paper, no. 32, Department of Industry, London.
Marshall, J.N., D. Bradley, C. Hodgson, N. Alderman and R. Richardson (2005), 'Relocation, relocation, relocation: assessing the case for public sector dispersal', *Regional Studies*, **39**, 767–87.
Meester, W.J. and P.H. Pellenbarg (2006), 'The spatial preference map of Dutch entrepreneurs: subjective rating of locations, 1983–2003', *Journal of Economic and Social Geography*, **94** (7), 364–76.
Moore, B. and J. Rhodes (1976), 'Evaluating the effects of British regional economic policy', *Economica*, **43**, 17–31.
Neary, J.P. (2001), 'Of hype and hyperbolas: introducing the economic geography', *Journal of Economic Literature*, **39**, 536–61.
Oosterhaven, J. (1981), *Interregional Input–Output Analysis and Dutch Regional Policy Problems*, Aldershot: Gower Publishing.
Oosterhaven, J. (1996), 'Dutch regional policy gets spatial', *Regional Studies*, **30** (5), 527–32.

Oosterhaven, J. (1997), 'Regional convergence and/or concentration: an overview', in Christen Sorensen (ed.), *Empirical Evidence of Regional Growth: The Centre–Periphery Discussion*, Copenhagen: Expert Committee to the Danish Ministry of the Interior, pp. 28–47.
Oosterhaven, J. and H. Folmer (1985), 'An interregional labour market model incorporating vacancy chains and social security', *Papers of the Regional Science Association*, **58**, 141–55.
Oosterhaven, J., F.G.J. Eding and D. Stelder (2001), 'Clusters, linkages and regional spillovers: methodology and policy implications for the two Dutch mainports and the rural north', *Regional Studies*, **35** (9) 809–22.
Ottaviano, G. and D. Puga (1998), 'Agglomeration in the global economy: a survey of the new economic geography', *World Economy*, **21**, 707–31.
Pellenbarg, P.H., L.J.G. van Wissen and J. van Dijk (2002), 'Firm migration', in P. McCann (ed.), *Industrial Location Economics*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, pp. 110–48.
Polèse, M. and R. Shearmur (2006), 'Why some regions will decline: a Canadian case study with thoughts on local development strategies', *Papers in Regional Science*, **85** (1), 23–46.
Porter, M. (1990), *The Competitive Advantage of Nations*, New York: Free Press.
Putnam, R.D. (1993), *Making Democracy Work: Civic Traditions in Modern Italy*, Princeton, NJ: Princeton University Press.
Richardson, H.W. (1979), 'Aggregate efficiency and interregional equity', in H. Folmer and J. Oosterhaven (eds), *Spatial Inequalities and Regional Development*, Boston and Dordrecht: Martinus Nijhoff-Kluwer, pp. 161–83.
Romer, P.M. (1986), 'Increasing returns and long-run economic growth', *Journal of Political Economy*, **94**, 1002–37.
Romer, P.M. (1990), 'Endogenous technological growth', *Journal of Political Economy*, **98** (5), 71–102.
SACTRA (1999), 'Transport and the economy', Standing Advisory Committee for Trunk Road Assessment, Department for Transport, London.
Saxenian, A.L. (2006), *The New Argonauts: Regional Advantage in a Global Economy*, Cambridge, MA: Harvard University Press.
Smith, V.L. (1979), *Research in Experimental Economics*, Vol. 1, Greenwich, CT: JAI Press.
Stilwell, F.J.B. (1972), *Regional Economic Policy*, London: Macmillan.
Stoffelsma, R.J. and J. Oosterhaven (1991), 'Social security and the income distribution: a spatio-temporal analysis for the Netherlands', *Public Finance*, **46** (2), 268–88.
Treyz, G.I. (1993), *Regional Economic Modeling: A Systematic Approach to Economic Forecasting and Policy Analysis*, Boston, MA: Kluwer.
Van Dijk, J. (1986), *Migratie en Arbeidsmarkt* (Migration and the Labour Market). Assen: Van Gorcum.
Van Dijk, J., H. Folmer, H.W. Herzog Jr. and A.M. Schlottmann (1989), 'Labor market institutions and the efficiency of interregional migration: a crossnation comparison', in J. van Dijk, H. Folmer, H.W. Herzog Jr. and A.M. Schlottmann (eds), *Migration and Labor Market Adjustment*, Dordrecht: Kluwer Academic Publishers, pp. 61–83.
Van Dijk, J., H. Folmer, H.W. Herzog Jr. and A.M. Schlottmann (2000), 'Worker endowments and the effects of institutions on earnings realization: a cross-nation comparison', *Journal of Regional Science*, **40** (1), 67–88.
Venables, A.J. (1996), 'Equilibrium locations of vertically linked industries', *International Economic Review*, **37**, 341–59.
Westlund, H. (2006), *Social Capital in the Knowledge Economy: Theory and Empirics*, Berlin, Heidelberg and New York: Springer.
Williamson, J.G. (1965), 'Regional inequalities and the process of national development', *Economic Development and Cultural Change*, **13**, 3–45.

24 New regional policies for less developed areas: the case of India

Maria Abreu and Maria Savona

24.1 Introduction

Regional disparities in growth and income levels present an important challenge for policy-makers in less developed countries, particularly in the context of increasing globalisation (World Bank, 2006). Globalisation heightens regional income disparities since wealthier regions typically have an infrastructure and skills advantage which, on top of wage differentials and less stringent environmental regulation, enables them to attract further domestic and foreign investment. A large number of recent empirical contributions have analysed the extent to which developing countries are able to benefit from trade liberalisation and other economic reform policies. However, only a few of these contributions are devoted to the impact of these policies on regional income disparities.[1]

This chapter reviews the empirical literature on regional policies in less developed countries, with an illustration based on the case of India. Our review shows that regional policies can complement or counteract the effects of national policies, with the effectiveness of specific regional policies depending on the degree of decentralisation of the policy-making process, the extent of sectoral specialisation across regions and the degree of regional variation in initial endowments in physical and social infrastructure.

We analyse this issue further in the context of regional policies pursued by the state governments of India over the period 1988–2001. India implemented national trade liberalisation policies starting in 1991, as part of a wider economic liberalisation strategy. Under the federal system of government in India, states have substantial jurisdiction over public health, transport infrastructure, labour regulation and education. We show that the policies pursued by individual states, particularly policies related to infrastructure, business regulation and education, have contributed greatly to the differential impact of national development policies across regions. The post-reform experience of India thus highlights the importance of formulating appropriate regional policies to complement national development strategies.

The remainder of the chapter is organised as follows. The next section presents an overview of the literature on regional effects of national policies, and the effectiveness of regional policies in promoting development. The following section discusses the issues further by presenting the case of India. We review the recent transformations of the Indian economy, justifying the reasons why India represents a particularly interesting case of a developing context in which regional disparities have been exacerbated by trade liberalisation policies. The section also reviews the policy interventions carried out at the regional level with the intention of counterbalancing the effects of trade liberalisation. Some empirical evidence on the patterns of regional growth before and after the introduction of trade liberalisation measures in the beginning of the 1990s is also presented. The last section summarises the main findings of the review and proposes a way ahead for

regional policies in less developed countries within the current context of increased globalisation.

24.2 Regional policies in less developed countries

Subnational policy-making has become an increasingly important topic in the development policy debate, particularly in relation to quality of governance and provision of public services (Bardhan, 2002). The World Bank, regional development agencies such as the Asian Development Bank and international and local non-governmental organisations (NGOs) have made decentralisation an important aspect of their policy agenda. While most of the policy debate has focused on deliberate or intended regional policies, for instance central government policies tailored to specific regions, or policies implemented by regional governments with some degree of decentralised power, there are also unintended regional policies or consequences that follow from the regional impact of central government policies such as trade liberalisation. The more spatially diverse the national economy, the greater the regional impact of national policies (Hill, 2000).

Unintended regional policies

The unintended consequences of national policies have been studied in the context of specific issues such as taxation, competition or trade liberalisation, although the new economic geography (NEG) literature has also analysed them in a general equilibrium framework (Baldwin et al., 2003; Ottaviano, 2003). The impact of unintended regional policies is particularly acute for less developed and transition countries because factor mobility is generally low, so that national policies have a larger impact on the more exposed regions, resulting in widening regional inequalities (Goldberg and Pavcnik, 2007).

Hill (2000) highlights the case of Indonesia, where import protection programmes for certain industries implicitly discriminate against the Outer Islands, because most of the protected industries are located in Java. Factor market regulations such as minimum wage regulations also have a larger effect on disadvantaged regions. The case of Mexico is also interesting because the national government started on a course of trade liberalisation since the mid-1980s which culminated with Mexico's accession to the North American Free Trade Agreement (NAFTA) in 1994. Sánchez-Reaza and Rodríguez-Pose (2002) explore the impact of trade liberalisation on regional inequality in Mexico, and find that increasing trade openness resulted in a shift from regional convergence to an increasing North–South divide, with regions close to the Mexico–US border benefiting the most from trade under the NAFTA agreement. The overall impact has been one of regional divergence. A similar pattern has been observed in China (Zhang and Zhang, 2003).

The impact of national policies on regional economies has also been studied in the context of monetary policy. Asymmetries in price transmission mechanisms can be large in countries where markets are less developed, and where there is less scope for arbitrage. Fielding and Shields (2006) find that changes in monetary policy lead to substantial and persistent changes in relative prices across South African provinces, pointing to a need for fiscal policies at the regional level to compensate for imbalances caused by national monetary targets.

We further deal with the issue of the unintended regional policies in section 24.3, which is devoted to the case of the regional effects of trade liberalisation policy in India.

Intended regional policies

The issues related to intended regional policies have been analysed in detail across a diverse range of disciplines, including economics, geography, sociology and political science (Bardhan, 2002). The study of regional policies is also closely related to the analysis of decentralisation, often referred to as 'fiscal federalism', particularly in the context of the United States. Fiscal federalism is a sub-field of public finance that studies the vertical structure of the public sector, and in particular, the optimal way of allocating responsibilities across different levels of government (Oates, 1999). A related literature within political science is 'multilevel governance', which studies the diffusion of authority across tiers of government, and has received a great deal of attention among European scholars (Hooghe and Marks, 2003). This literature has argued that jurisdictions can be designed around particular communities, such as districts, counties or provinces, or around particular problems, such as providing a particular local service or solving a common resource problem. It is also important to note that decentralisation can take many forms, among them political, administrative and fiscal. In addition, the degree of autonomy experienced by the decentralised jurisdictions may differ considerably from that accorded to them by the law (Treisman, 2002).

However, the conclusions of the fiscal decentralisation literature do not immediately carry over to less developed countries (Bardhan, 2002). The fiscal federalism literature is to a large extent based on the assumption that individuals are mobile between jurisdictions, and will seek out the most attractive combination of local public goods and taxes. However, mobility in less developed countries is often fairly limited, particularly in large countries with substantial cultural heterogeneity. In addition, accountability and governmental transparency problems are more severe in developing countries, and a lack of qualified staff may limit the quality of local governance. Local tax-setting powers that provide part-funding for local projects may also not be feasible for low-income jurisdictions, and taxing immobile factors such as land is often politically infeasible.

The literature on regional policy-making in less developed countries has therefore focused on issues of accountability, governance and achieving a balance between local fiscal discipline and fiscal equalisation across regions (Bardhan, 2002). Just as in the decentralisation literature in high-income countries, an important issue is at which level of government policies should be set and decisions made, which in turn depends on the type of service or public good to be provided. The literature on fiscal decentralisation, for instance, argues that: 'the provision of public services should be located at the lowest level of government encompassing, in a spatial sense, the relevant benefits and costs' (Oates, 1999, p. 1122), but the political economy of decentralisation in developing countries implies that in some instances the optimum level of decentralisation may be more dispersed than is economically optimal. Matters are often confounded by the administrative design of regional boundaries in ways that are not compatible with economic and political principles.

Revenue sharing and quality of governance

The financing mechanisms for fiscal decentralisation can take different forms, among them local self-finance (through local taxes), conditional and unconditional transfers from the central government. The fiscal federalism literature has argued that conditional transfers in the form of matching grants (where the central government finances a specific share of the local government's expenditure) are suitable in cases where the provision of

local services generates benefits to residents in other jurisdictions, while unconditional transfers are the appropriate vehicle for fiscal equalisation (Oates, 1999). However, self-finance and competition for mobile factors across jurisdictions may result in a 'race to the bottom', where jurisdictions compete to offer low taxes and lax regulatory environments, and provide suboptimal amounts of public services (Cai and Treisman, 2004). Possible solutions include the charging of user fees for specific services and the centralisation of some welfare provision programmes (Oates, 1999).

An important issue, which is particularly relevant in the context of developing countries, is whether fiscal decentralisation can help to improve the quality and accountability of governments. The evidence is fairly mixed, and depends on specific legal and political contexts. In some instances fiscal decentralisation can lead to increased corruption as local government is captured by the elite (Bardhan and Mookherjee, 2000), while introducing new tiers of government may lead to increased corruption if the tiers do not collude (Shleifer and Vishny, 1993). In a cross-country study of fiscal decentralisation and corruption, Fisman and Gatti (2002) find that fiscal decentralisation results in lower levels of corruption, while Treisman (2002) argues that the effect varies with the type of decentralisation process. In particular, he finds that countries with a large number of tiers tend to have higher perceived levels of corruption. In practice, the effects of decentralisation on corruption and quality of governance are likely to be highly context-specific.

Accountability can also be strengthened by 'yardstick competition', where jurisdictions are compared in terms of performance (Besley and Case, 1995). This allows voters to judge the efforts of local public officials, where their competence cannot be directly observed. Yardstick competition can also be used in policy experiments, where reforms are introduced in some localities and not in others. This approach was used with some success in the early period of the market reforms in China (Maskin et al., 2000).

Infrastructure, irrigation and utilities

The fiscal federalism literature has generally argued that decentralisation is to be preferred when tastes are heterogeneous and there are no spillovers across jurisdictions (Oates, 1972). The traditional view is thus that infrastructure that results in spillovers across jurisdictions should be implemented by a higher tier of government, while infrastructure that affects only residents in the jurisdiction should be implemented by local governments. However, infrastructure development is a major expenditure outlay, and may not be within the reach of local governments. In addition, subsidies and failure to implement user fees can lead to severe fiscal problems, as is, for instance, the case with the Indian state electricity boards (Ahluwalia, 2002b). The literature also suggests that quality control and oversight should remain the remit of the central government, to ensure that infrastructure standards are consistent across jurisdictions and to support local governments with technical resources if there is a lack of local qualified personnel.

Infrastructure provision can result in both increases and decreases in interregional inequality, depending on whether infrastructure investments are spread uniformly across jurisdictions and whether the type of investment is targeted to local needs. For instance, in a survey of regional policies in South-East Asia, Hill (2000) finds that countries such as Malaysia and Indonesia, which implemented fiscal equalisation programmes and invested heavily in physical infrastructure, experienced less regional divergence than other countries in the region.

Case study evidence has also revealed that fiscal and political decentralisation often leads to a better targeting of infrastructure investments to local needs. In Bolivia, a decentralisation initiative started in 1994 which allocates a greater share of national tax revenues and devolves political power to the municipalities resulted in a large shift of public resources from large-scale production to spending on services such as water and sanitation, a process that was mainly driven by expenditure shifts in poor municipalities (Faguet, 2004). A similar process was observed in a survey of villages in India (Foster and Rosenzweig, 2001). More generally, a study of decentralisation and infrastructure provision in the *World Development Report 1994* (World Bank, 1994) concluded that the maintenance of road and water supply systems was higher in cases where local communities were given part of the management responsibility of the projects.

Welfare, education and health

The fiscal federalism literature has argued that welfare and poverty alleviation programmes should ideally be administered by the central government, because the mobility of individuals across jurisdictions may result in policies that provide too little assistance to the poor, also known as the 'race to the bottom'. This has in part been the experience of developed countries such as the United States (Oates, 1999). However, given the information asymmetries present in many developing countries, a case can be made for the targeting of national welfare programmes to the poor by local administrations. An example of the latter is a central-government food-for-education programme in Bangladesh which, although managed by the central government, was based on targeting of households within communities by local school management committees. Galasso and Ravallion (2001) find that the targeting of the programme was mildly pro-poor. In a study of a targeted social assistance programme in Albania, Alderman (1998) finds that the efficiency of targeting and cost-effectiveness of the programme improved after it was decentralised in 1995.

In recent years there has been a global tendency towards decentralisation in education. The education literature has argued that decentralisation can help to improve efficiency, transparency and accountability, encourage participation and improve coverage. However, the process has proved to be complex, because of the number of complementary functions involved such as curriculum design, student evaluation, textbook production and distribution, teacher recruitment and pay, construction of schools, and so on.

The empirical literature on the impact of decentralisation on educational outcomes and school participation has generally been positive. Moves to transfer management tasks to local councils involving parents in Nicaragua have had a positive impact on student performance (King and Özler, 1998), while studies of the community-administered EDUCO programme in El Salvador have found relatively little impact on student achievement but positive effects on student and teacher absenteeism (Jimenez and Sawada, 1998). The experience of education decentralisation in Chile and Brazil suggests that it can lead to improvements in participation, but the process is not sufficient to reduce inequalities in education outcomes across jurisdictions (Winkler and Gershberg, 2000). Similar results were found in impact evaluations of education decentralisation in Zimbabwe and New Zealand (Fiske, 1996).

The arguments for decentralisation of healthcare are significant, and include a better targeting to local needs, with resulting improvements in efficiency and cost-effectiveness,

greater involvement of local communities, and improved coordination with other local services. The empirical evidence, however, is mixed, and there are relatively few comprehensive studies of the impact of regional health policies on quality and coverage of healthcare. The issues are further compounded by the political nature of some health-related services, such as family planning.

Past experience shows that the impact of health decentralisation depends heavily on its design, and in particular on the quality of the coordinating mechanism across jurisdictions and with the central government, the provision of adequate financing and the good management of capacity constraints and staff shortages (Mills et al., 1990).

24.3 The case of India

India has a heritage of democratic institutions and a highly developed system of education as compared to other developing countries, which facilitates its potential for catching-up. Particularly interesting as well is its regional map, not only in terms of sectoral structure, growth patterns and per capita wealth, but also in terms of cultural and religious differences. The sectoral specialisation pattern across Indian states is strikingly varied. India is the new global pole for international business process outsourcing (BPO) of information technology (IT) and business services, particularly in the coastal states, which contrasts with a large presence of farming activities and rural poverty in many of its hinterland states. India is therefore an interesting case to study the regional effects of a radical shift of national policy towards international trade liberalisation, which affects only a limited number of sectors. It is important to disentangle the role of factors such as the international environment, sectoral structural change, information and communication technology (ICT) related technological change, from the direct effect of national policies, which have paved the path of India's process of catching-up. This is what we aim to do in this section.

From independence to the present

Since India's independence in 1947, several scholars have looked at the history of the Indian economy (among others, Bhagwati and Desai, 1970; Joshi and Little, 1996). The global attention is now focused on India's pattern of development and reform process, which experienced a substantial shift in the 1980s and 1990s (Rodrik and Subramanian, 2004; Banerjee and Iyer, 2005; Kochhar et al., 2006; Aghion et al., 2006). However, the aggregate and regional impact of the national reforms of the 1990s is still a matter for debate.

The gross domestic product (GDP) growth rate of the first three decades after independence – of a magnitude of 3.5 per cent per year – has been labelled the 'Hindu' rate of growth, in the sense that it contrasts negatively with the higher rates of other Asian, particularly East Asian, countries (Rodrik and Subramanian, 2004). The 1980s and 1990s, however, saw a boost to a 6–8 per cent annual GDP growth. The challenge is therefore to identify the main drivers of the 'Indian miracle' and to distinguish the type of policies that ignited the transition to high growth in the 1980s and early 1990s from those which have been put forward by the government to maintain the high rate of growth over the late 1990s and early 2000s.

The widespread view on the determinants of the 'Indian miracle' has attached great importance to the processes of liberalisation started in 1991. In the wake of a severe

Table 24.1 Output growth in the Indian economy, annual change (%)

	1988	1990	1995	2001	2002	2003	2004	2005
Agriculture	15.5	4.1	−0.9	6.2	−6.9	10.0	0.7	3.9
Industry	9.1	7.7	11.6	2.7	7.0	7.6	8.6	8.7
Services	7.3	6.8	10.5	7.1	7.3	8.2	9.9	10.0
GDP	10.5	5.6	7.3	5.8	3.8	8.5	7.5	8.4

Source: Asian Development Bank (2006).

balance-of-payments crisis, the Indian government introduced in 1991 a number of wide-ranging economic reforms. The reforms included trade liberalisation, opening up the economy to foreign investment, and a reduced role for the state in economic activity. There was also a shift towards the decentralisation of government, which was formalised through constitutional amendments in 1992 (Bhide and Shand, 2003).

The reform process initially focused on industrial and trade policy, and in particular, on reducing state controls on private investment. Prior to 1991 private enterprises faced a number of restrictions; the scale and location of enterprises was often determined by the state, and private ownership was curtailed in a large number of sectors. Local industries were also protected by high tariffs and import restrictions. The economic reforms of the 1990s sought to phase out most of the import licensing restrictions, lower import duties and remove state controls on private enterprises.

The reforms opened up a large number of sectors to private investment, including mining and quarrying, basic metals, oil refining and exploration, road and air transport, telecommunications, pharmaceuticals and the financial sector. Industrial policy prior to 1991 restricted the scale of enterprises in several manufacturing sectors to small-scale operations. Although this restriction was gradually lifted for a few sectors, notably garments and textiles, it initially implied that the reforms were more conducive to growth in the services sector, while the manufacturing firms continued to be affected by stiff competition from South-East Asian economies, notably from China (Ahluwalia, 2002b).

The reform process has led to a revolution in the services sector, which is now India's fastest-growing and most dynamic sector (Table 24.1). The Indian services boom has been led by software development and business process outsourcing (BPO) firms, offering services ranging from answering the telephone in call centres to writing software code, filing tax returns, processing insurance claims, preparing presentations and undertaking market research. The expansion in BPO services has drawn on India's comparative advantage in terms of a highly qualified and English-speaking workforce.

The reforms have had less of an impact on agriculture, a sector which provides employment and sustenance for a large proportion of the Indian population (Harriss-White, 2003). Import tariffs and duties on agricultural products have largely remained in place, although the agricultural exports have risen. The process of decentralisation which accompanied the reforms led to a worsening of the government's fiscal position, particularly of state finances, with a resulting decline in investment in areas that are critical for agricultural productivity, such as irrigation, water management and rural infrastructure (Ahluwalia, 2002b).

A different view on the characteristics and impact of the 'Indian miracle' comes from scholars such as Rodrik and Subramanian (2004), who argue that the catching-up process labelled as the 'Indian miracle' started a full decade before 1991. The authors argue that India's economic boost should be partly attributed to an 'attitudinal shift' of the national government in the beginning of the 1980s, which has favoured a pro-business set of measures rather than a pro-market one. The latter tend to favour new entrants and consumers by removing impediments to market and typically pave the path for full trade liberalisation. The former are instead a set of measures such as reduction of corporate taxes and price controls which favour incumbents and existing business. Among other factors such as the above-mentioned remarkable heritage of democratic institutions and high-level education system, the fertile environment created in the 1980s is mostly responsible for the 'Indian miracle', as the pro-business approach 'had the political economy merit of avoiding the creation of losers' (Rodrik and Subramanian, 2004, p. 38).

Kochhar et al. (2006) offer a rather different sectoral and regional perspective of the history of the catching-up process in India. They argue that the idiosyncratic policy interventions since independence and prior to the 1980s have favoured a far more diversified presence of skill-intensive and high labour-productive manufacturing industries than in other developing countries. Interestingly enough, the presence of private business services has been relatively low up to the beginning of the 1980s, as compared to the share of skill-intensive segments in the public sector, such as telecommunications. During the 1980s, the peculiarity of a skill-intensive growth instead favoured private services, including business services (also outsourced by developed countries), and information technology as well as finance.

Between 1980 and 2002, India's share of services in value-added has risen from 37 per cent to 49 per cent, a striking figure for a developing country. The share of manufacturing activities remained unchanged at 16 per cent. All in all, India has therefore experienced a very peculiar pattern of structural change, as compared to its neighbouring countries such as China, which drew employment from agriculture to manufacturing. Rather, India boosted its pool of high-skilled employment through remarkably high investments in tertiary education for its level of per capita income, which has enabled a generation of skills-intensive service industries (Kochhar et al., 2006).

Tightly linked with the issue of changes in the sectoral structure of the Indian economy and the 1990s reforms, is the assessment of the pro-poor[2] impact of the 'Indian miracle' since the 1980s. Datt and Ravallion (2002) and Ravallion and Datt (2002) identify a clear shift in the poverty-reduction effect of patterns of growth before and after the 1990s. In particular, they argue that the available statistical evidence does not show an increase in overall inequality prior to the 1990s. Rather, the figures on the elasticity of national poverty measures to economic growth between 1958 and 1991 show that the individuals well below or near the poverty lines benefited from economic growth.

However, the trend reversed in the 1990s, as also shown in Bhalla (2000). Datt and Ravallion (2002) and Ravallion and Datt (2002) argue that the nature of the growth process characterising India overall – and Indian states in particular – has been such that the pro-poor effect has been substantially lower (when not negative) in the 1990s compared to the previous period.

Consistent with the picture drawn by Rodrik and Subramanian (2004) and Kochhar et al. (2006), Ravallion and Datt (2002) and Datt and Ravallion (2002) find large

Table 24.2 Growth of state domestic product (SDP) by sector

	SDP growth 1991–1993	SDP growth 1994–2004	Sectoral SDP growth 1994–2004		
			Agriculture	Industry	Services
Maharashtra	8.6	5.4	1.1	2.8	7.4
Delhi	8.5	13.8	3.0	9.3	16.8
Karnataka	7.5	6.8	0.8	7.1	10.6
Gujarat	7.0	8.1	10.9	8.2	8.2
Kerala	6.5	5.3	−0.3	5.5	8.3
West Bengal	6.0	7.4	3.7	5.7	9.6
Orissa	5.8	4.2	1.7	6.6	6.4
Tamil Nadu	5.5	5.2	0.2	2.7	8.5
Andhra Pradesh	4.8	5.9	3.4	6.4	7.8
Punjab	4.5	4.5	2.3	3.5	6.9
Madhya Pradesh	3.6	5.0	3.4	5.3	6.6
Haryana	2.2	6.4	2.4	7.0	9.2
Uttar Pradesh	1.3	3.8	2.7	3.3	5.6
Rajasthan	−0.3	7.2	8.0	8.3	7.5
Bihar	−3.1	4.9	5.0	7.7	6.6

Source: Reserve Bank of India (2006).

cross-state differences in the elasticity of poverty to the growth of non-farm activities. This has brought about a much diversified cross-state pro-poor impact of growth, depending on the initial level of rural development, farm productivity and human capital development. In particular, states with a higher poverty elasticity did not experience high rates of non-farm growth, to the point that, 'if not for the sectoral and geographic imbalance of growth . . . the national rate of growth would have generated a rate of poverty reduction that was double India's historical trend rate' (Datt and Ravallion, 2002, p. 106).

Overall, taking into account sectoral and regional specificities, and the initial level of rural and human capital development, the effect of national policies might be detrimental to a balanced cross-state development. We further develop this issue in the next section.

Regional aspects of the reform process

The gradual nature of the reform process and the differences in sectoral mix across states mean that its effects have been felt more keenly by some states than by others. India is a federal union of 28 states and seven federally governed territories, including the National Capital Territory of Delhi. Coastal states and territories such as Maharashtra, Delhi, Karnataka and Gujarat benefited the most in the initial 1991–93 period of the reform process (Table 24.2) and continued to grow strongly in the 1994–2004 period, while the growth rates of the mainly agricultural hinterland states Bihar and Uttar Pradesh have remained low. The states of West Bengal, Rajasthan and Haryana experienced intermediate levels of gross state product (GSP) growth in the 1991–93 period, but have caught up in recent years.

The process of trade liberalisation has resulted in a reallocation of economic activity towards areas of higher comparative advantage, mostly in the coastal regions. Growth in the services sector has in recent years been particularly high in Karnataka, whose capital Bangalore is a major IT hub, as well as Delhi, West Bengal and Haryana.

It is still debatable whether these idiosyncratic changes in the economic structure of India towards services have had a positive aggregate impact on its catching-up process (Nayyar, 1988; Dasgupta and Singh, 2005). Nonetheless, it is often argued that this sectoral structural change has exacerbated cross-state disparities, particularly between fast-growing coastal states and slow-moving hinterland ones, still penalised by a poor set of infrastructure and institutions. Kochhar et al. (2006) argue that, as a consequence, most business-friendly peninsular states' specialisation patterns will lead them to developed countries' growth rates. Conversely, it would be consistent for the laggard hinterland states, like other traditional patterns of transition from underdevelopment, to shift towards labour-intensive manufacturing sectors, with the support of specifically tailored state policies, as we will see in the next section.

In line with this picture, Sachs et al. (2002) also find evidence in support of a growing cross-state growth divergence in India. In particular, it is claimed that the opening-up to foreign trade has created further disparities between coastal and inland laggard states. Even with higher overall growth, this process is likely to create and/or worsen cross-state migration and deepen the cycle of cross-state inequality.

The differential impact of the reforms has also been a function of variations in the business climate across states that go beyond geographical location. States with high levels of human capital, good-quality infrastructure and low levels of state government regulations have benefited the most. An investment climate assessment of India carried out by the Confederation of Indian Industry (CII) in 2003 in collaboration with the World Bank found significant variations in the perception of obstacles to business and growth across Indian states (Table 24.3). The pattern of responses indicates that six states which attracted the largest share of domestic and foreign direct investment (FDI) were also identified as having the better investment climates by respondents (World Bank, 2004).

The differential growth process across states has led to a worsening regional inequality, resulting from an increased dispersion in the growth rates of individual states (Ahluwalia, 2002a, 2002b; Baddeley et al., 2006). There is a strong correlation between the growth performance of the states and their human development indicators (HDIs) (Table 24.4). While poverty has generally fallen across all states since 1991, it has fallen by a significantly larger amount in fast-growing states such as Kerala than in slower-growing states such as Bihar and Uttar Pradesh (Datt and Ravallion, 2002). A possible explanation is that differences in initial conditions such as a lack of access to credit, unequal distribution of assets such as land and low levels of education limit the ability of the poor to benefit from the reform process. Since most of these factors are within the policy remit of the individual states, the state governments have a major role to play in ensuring that central government reforms have a positive impact on human development in their regions.

Regional development policies: lessons from India

Under the federal system in India, the states play a key role in formulating policy. The state governments have jurisdiction over road transportation (except national highways),

Table 24.3 Business climate indicators: percentage of firms identifying an issue as an obstacle to growth

	Labour regulation	Infrastructure	Power supply	Skill shortages	Corruption
Maharashtra	18.9	31.6	22.3	22.3	9.3
Delhi	68.5	4.8	1.2	1.2	8.3
Karnataka	51.8	61.2	59.3	59.3	23.1
Gujarat	7.0	24.4	5.7	5.7	13.7
Kerala	5.9	19.8	5.9	5.9	2.0
West Bengal	9.1	31.7	5.2	5.2	19.9
Tamil Nadu	30.3	40.9	46.3	46.3	18.2
Andhra Pradesh	2.4	17.3	41.1	41.1	11.9
Punjab	2.8	11.7	3.9	3.9	10.0
Madhya Pradesh	2.1	49.3	54.0	54.0	3.6
Haryana	26.5	30.7	27.1	27.1	5.9
Uttar Pradesh	2.3	51.7	38.1	38.1	9.8

Source: World Bank (2004).

Table 24.4 Human development indicators by state, 1999–2002

	Rural poverty	Urban poverty	Literacy rate	Infant mortality rate (per 1000)	Safe drinking water (% households)
Maharashtra	23.7	26.8	76.9	45	79.8
Delhi	0.4	9.4	81.7	33	97.2
Karnataka	17.4	25.3	66.6	55	84.6
Gujarat	13.2	15.6	69.1	60	84.1
Kerala	9.4	20.3	90.9	10	23.4
West Bengal	31.9	14.9	68.6	49	88.5
Orissa	48.0	42.8	63.1	87	64.2
Tamil Nadu	20.6	22.1	73.5	44	85.6
Andhra Pradesh	11.1	26.6	60.5	62	80.1
Punjab	6.4	5.8	69.7	51	97.6
Madhya Pradesh	37.1	38.4	63.7	85	68.4
Haryana	8.3	10.0	67.9	62	86.1
Uttar Pradesh	31.2	30.9	56.3	80	87.8
Rajasthan	13.7	19.9	60.4	78	68.2
Bihar	44.3	32.9	47.0	61	86.6

Note: The statistics in columns 1 and 2 refer to the period 1999–2000; columns 3 and 5 to 2001; columns 4 and 5 to 2002.

Source: National Human Development Report 2001, United Nations Development Programme (UNDP) and Economic Survey 2005–06, Government of India (2002) and Ministry of Finance (2006).

agriculture, land rights, irrigation, water supply, public health and urban services. The state and the national government have joint responsibility over a number of areas including education, trade unions and labour, and the provision of electricity (these powers are known as the concurrent list). Both the national government and the states can legislate in these areas and the state laws must be consistent with central laws, but in practice most of the delivery falls to the states.

A major shortcoming of the reform process has been the poor level of public finances, both at the central government level and the state level. Overall revenue deficits have increased significantly since the late 1980s, while in previous years the balance had been positive or close to zero for most states (Singh, 2006). The issue of poor state finances adds to the difficulties of state-level policies, by restricting the options available to the state governments. Reform of the tax system is a priority, but is compounded by the fact that the state governments legislate on sales and agricultural taxes, while the central government controls income taxes. A coordinated approach is essential, but has proven difficult to implement in practice.

An important drain of state finances is caused by the State Electricity Boards (SEBs), which have sole responsibility for generating, transmitting and distributing power within the states. The financial position of the SEBs has been affected in recent years by the policy of subsidised tariffs for farmers and household consumers. The subsidies are partly financed through high tariffs levied on industrial and commercial users. The resulting deficits have led to power shortages and unreliable supplies (Ahluwalia, 2002a). Lack of a reliable supply of electricity was one of the major issues cited by respondents of the Investment Climate Assessment (World Bank, 2004) as a major impediment to business growth (Table 24.3). The state of Orissa has pioneered reforms in this area, establishing a commission to review existing tariffs and restructure the local SEB (Baddeley et al., 2006).

The poor condition of physical infrastructure in several states is another major impediment to growth, particularly for the manufacturing sector. Several states, notably Karnataka, have recognised the importance of improving their infrastructure in order to attract industry, although inevitably it is mostly the states that are already growing and have access to greater public resources. Telecommunications is an area where the reforms have succeeded, a factor that has greatly contributed to the growth of the information technology (IT) sector. Some progress has been made in ports and civil aviation, which fall under the remit of the central government. However, roads and railways remain poor, further hindering the manufacturing and agricultural sectors, particularly in the northern states.

High levels of basic and advanced education and a good social infrastructure have been found to be crucial in ensuring that the national economic reforms benefit the poor (Datt and Ravallion, 2002). Skill shortages have also been identified by respondents as a major impediment to business growth, particularly in the service-oriented states. A significant amount of variation in human capital indicators across states can be attributed to differences in underlying social and demographic features, such as gender and caste structures (Baddeley et al., 2006). In this context Kerala deserves a special mention; the state is unusual in that it has long had high levels of education and other human development indicators, which are considerably higher than the Indian average. For instance, the literacy rate in Kerala is almost 91 per cent, compared to an Indian average of 65 per cent. However, Kerala has to a large extent failed to capitalise on its schooling advantage, partly

through its chronic political instability. Nonetheless, its growth rate has picked up in recent years, through remittances from workers who have migrated to other Indian states and to the Middle East, and a boom in the tourism industry.

An important issue affecting all states is the need for political stability and good governance. An indirect consequence of the reform process implemented in 1991 has been a change in the balance of power of the union, favouring the states over the central government. The result has been an increasing shift to regional politics, with regional political parties dominating over national parties. Combined with the strong tendency in Indian politics to favour new candidates over incumbents, this has led to a shift in focus towards short-term policy goals. The evidence suggests that an important aspect of state-level reforms in India will be to ensure transparent governance, which must go hand-in-hand with state-level fiscal reforms.

Much emphasis has recently been put by the international institutions such as the World Bank and the International Monetary Fund (IMF) on the concept of 'good governance' at the state level. As pointed out by Nayyar (2002, 2007), good governance is an essential ingredient in the process of development at both the national and state level. However, it is also true that the Indian states which have experienced the fastest growth rates – that is, the six north-east heartland ones – also have the worst quality of governance in terms of the IMF and World Bank parameters, whereas the case of Kerala is, on the contrary, an example of good governance – in terms of literacy rates, gender equality, level of overall HDI – and yet slow economic growth.

According to Nayyar (2002, 2007), different degrees and types of coordination between state governance and national policy would entail different outcomes. Firstly, the state governance might be able to introduce adequate correctives to a national policy which, as in the case of trade liberalisation in India, has exacerbated cross-state disparities. Secondly, the state governance might be able to reinforce the positive effects of national policies by supporting and facilitating them. Thirdly – a case often occurring in large less developed countries such as India – potentially good national policies are offset by bad governance at the state level.

The experience of India has shown that industrial liberalisation by the central government must be accompanied by appropriate state policies, as in the first two cases mentioned above. It is crucial to ensure a full and equally distributed participation of the states in the design of national policy. Applying equal rules – that is, national policies – to unequal 'participants' to the game results in getting unequal outcomes. These latter have to be counterbalanced with measures taken at a different level of policy decision – that is, state policy. This has to be achieved in turn by ensuring an equally distributed participation of people, which are the ultimate actors of what has been labelled 'good governance'.

24.4 Conclusions

The chapter has reviewed the regional effects of national policies in a developing-country context. In many less developed countries, both low- and middle-income countries, policies of economic and trade liberalisation have been implemented, which have had an uneven impact at the regional level. This has represented our point of departure.

In section 24.2 we have reviewed the literature dealing with the regional policy tools available in a developing context involved in a process of catching-up. In section 24.3 we have focused on the case of India and tried to disentangle the reasons behind the histories

of success and/or failure of state policies implemented during a period of radical changes in government trade and foreign policy. These latter are thought to have ignited the process of catching-up which has generally characterised its neighbouring countries, such as Korea and China.

India is a particularly interesting case since it is characterised by large cross-state differences in terms of cultural and religious practices, and of initial endowments of resources and skills. The change in the sectoral economic structure across states which has accompanied the overall 'Indian miracle' has been idiosyncratic to India and unusual with respect to other developing countries. None of its neighbouring countries – and none of the Western developed countries if we consider the process of development that has occurred since the Industrial Revolution – has experienced such a radical change in the economic structure from farming activities to services. India is now one of the global poles for business process outsourcing for IT-related and business services.

However, we have shown that the benefit of the 'Indian miracle' has not been equally spread across states. First and foremost, the reasons for this can be found among the wide cross-state disparities in initial levels of endowments and skills, as pointed out by Datt and Ravallion (2002). These disparities have both caused and been exacerbated by the changes in the sectoral structure across states, which have followed an unusual pattern, driving fast-growing regions towards skill- rather than labour-intensive industries.

Both initial disparities and sectoral structural changes have caused the impact of the 1991 reforms to be widely different across states. Even the pro-poor elasticity of economic growth after 1991 has fallen on average, unlike in previous periods.

The position of regional and international lending associations such as the Asian Development Bank, the International Monetary Fund and the World Bank has been to focus on the priority of loosening the rigidity of the labour market – to favour hinterland states with an unused pool of human resources for labour-intensive manufacturing activities – and improving state governance, as proposed in the seminal paper by Kaufmann et al. (1999).

However, as pointed out by Datt and Ravallion (2002) and Nayyar (2002, 2007), to ensure a more egalitarian impact of growth (especially in non-farm activities, as is the case, due to the peculiar pattern of sectoral structural change in India) regional policies should also include on improvement to rural and human resource development and a more egalitarian distribution of land. This is all the more crucial, as we have seen in sections 24.3, as the reform process started in 1991 has had a more favourable impact on the fast-growing coastal states, with better physical infrastructure and human development indicators.

Notes

1. For an extensive empirical review, see Cornia (2003, 2004), Cornia and Rosignoli (2006). For analysis of increased income inequality in India see Jha (2000) and Deaton and Drèse (2002).
2. The effect of economic growth on the poverty reduction is often labelled as the 'pro-poor' impact of economic growth.

References

Aghion, P., R. Burgess, S. Redding and F. Zilibotti (2006), 'The unequal effects of liberalisation: evidence from dismantling the Licence Raj in India', Centre for Economic Performance, London School of Economics, CEP Discussion Paper n. 728.

Ahluwalia, M. (2002a), 'State level performance under economic reforms in India', in A. Krueger (ed.), *Economic Policy Reforms and the Indian Economy*, Chicago, IL: University of Chicago Press, pp. 91–122.
Ahluwalia, M. (2002b), 'Economic reforms in India since 1991: has gradualism worked?', *Journal of Economic Perspectives*, **16** (3), 67–88.
Alderman, H. (1998), 'Social assistance in Albania: dcentralization and targeted transfers', LSMS Working Paper No. 134, World Bank.
Asian Development Bank (2006), *Key Indicators of Developing Asian and Pacific Countries 2006*, Manila, Philippines.
Baddeley, M., K. McNay and R. Cassen (2006), 'Divergence in India: income differentials at the state level, 1970–97', *Journal of Development Studies*, **42** (6), 1000–1022.
Baldwin, R., R. Forslid, P. Martin, G. Ottaviano and F. Robert-Nicoud (2003), *Economic Geography and Public Policy*, Princeton, NJ: Princeton University Press.
Banerjee, A. and L. Iyer (2005), 'History, institutions and economic performance: the legacy of colonial land tenure system in India', *American Economic Review*, **95** (4), 1190–1213.
Bardhan, P. (2002), 'Decentralization of governance and development', *Journal of Economic Perspectives*, **16** (4), 185–205.
Bardhan, P. and D. Mookherjee (2000), 'Capture and governance at local and national levels', *American Economic Review*, **90** (2), 135–9.
Besley, T. and S. Case, (1995), 'Incumbent behavior: vote-seeking, tax setting and yardstick competition', *American Economic Review*, **85** (1), 25–45.
Bhalla, S. (2000), 'Growth and poverty in India: myth and reality', http://www.oxusresearch.com.economic.asp.
Bhagwati, J. and P. Desai (1970), *Planning for Industrialisation*, New Delhi: Oxford University Press.
Bhide, S. and R. Shand (2003), 'Growth in India's state economies before and with reforms: shares and determinants', in R. Jha (ed.), *Indian Economic Reforms*, Basingstoke: Palgrave Macmillan.
Cai, H. and D. Treisman (2004), 'State corroding federalism', *Journal of Public Economics*, **88** (3–4), 819–43.
Cornia, G.A. (2003), 'Changes in the distribution of income over the last two decades: extent, sources and possible causes', paper presented at the 2003 Annual Meeting of the Società Italiana degli Economisti, Salerno, October.
Cornia, G.A. (2004), *Inequality, Growth and Poverty in an Era of Liberalisation and Globalisation*, Oxford: Oxford University Press.
Cornia, G.A. and S. Rosignoli (2006), 'Trends in the distribution of income over the period 1950–2004', University of Florence and IRPET, mimeo.
Dasgupta, S. and A. Singh (2005), 'Will services be the new engine of economic growth in India?', University of Cambridge, CBR Working Paper no. 308.
Datt, G. and M. Ravallion (2002), 'Is India's economic growth leaving the poor behind?', *Journal of Economic Perspectives*, **16** (3), 89–108.
Deaton, A. and J. Drèse (2002), 'Poverty and inequality in India: a re-examination', Centre for Development Economics, Working Paper no. 107.
Faguet, J.P. (2004), 'Does decentralization increase government responsiveness to local needs? Evidence from Bolivia', *Journal of Public Economics*, **88** (3–4), 867–93.
Fielding, D. and K. Shields (2006), 'Regional asymmetries in monetary transmission: the case of South Africa', *Journal of Policy Modeling*, **28** (9), 965–79.
Fiske, E.B. (1996), *Decentralization of Education: Politics and Consensus*, Washington, DC: World Bank.
Fisman, R. and R. Gatti (2002), 'Decentralization and corruption: evidence across countries', *Journal of Public Economics*, **83** (3), 325–45.
Foster, A. and M. Rosenzweig (2001), 'Democratization, decentralization and the distribution of local public goods in a poor rural economy', mimeo, University of Pennsylvania.
Galasso, E. and M. Ravallion (2001), 'Decentralized targeting of an anti-poverty program', Development Research Group Working Paper, World Bank.
Goldberg, P. and N. Pavcnik (2007), 'Distributional effects of globalization in developing countries', *Journal of Economic Literature*, **45** (1), 39–82.
Government of India (2002), *National Human Development Report of 2001*, New Delhi: Oxford University Press.
Harriss-White, B. (2003), *India Working: Essays on Society and Economy*, Cambridge: Cambridge University Press.
Hill, H. (2000), *Intra-Country Regional Disparities*, paper presented at the Second Asia Development Forum, 5–8 June, Singapore.
Hooghe, L. and G. Marks (2003), 'Unraveling the central state, but how? Types of multi-level governance', *American Political Science Review*, **97** (2), 233–43.
Jha, R. (2000), 'Reducing poverty and inequality in India: has liberalisation helped?', UNU/WIDER Working Paper no. 207, UNU/WIDER, Helsinki.

Jimenez, E. and Y. Sawada (1998), 'Do community-managed schools work? An evaluation of El Salvador's EDUCO program', Working Paper Series on Impact Evaluation of Education Reforms, No. 8, World Bank.
Joshi, V. and I.M.D. Little (1996), *India's Economic Reforms, 1991–2001*, Oxford: Clarendon Press.
Kaufmann, D., A. Kraai and P. Zoido-Lobatón (1999), 'Governance matters', World Bank Policy Research Working Paper 2196.
King, E. and B. Özler (1998), 'What's decentralization got to do with learning? The case of Nicaragua's school autonomy reform', Development Research Group Working Paper, World Bank.
Kochhar, K., U. Kumar, R. Rajan, A. Subramanian and I. Tokatlidis (2006), 'India's patterns of development: what happened, what follows?', IMF Working Paper 06/22.
Maskin, E., Y. Qian and C. Xu (2000), 'Incentives, information and organizational form', *Review of Economic Studies*, **67** (2), 359–78.
Mills, A., J. Vaughan, D. Smith and I. Tabibzadeh (1990), 'Health system decentralization: concepts, issues and country experience', World Health Organization, Geneva.
Ministry of Finance (2006), *Economic Survey 2005–2006*, New Delhi, India.
Nayyar, D. (1988), 'The political economy of international trade in services', *Cambridge Journal of Economics*, **12** (2), 279–98.
Nayyar, D. (2002), *Governing Globalisation*, Oxford: Oxford University Press.
Nayyar, D. (2007), 'Development in a world of globalisation: redesigning strategies and rethinking development', lecture held at the Cambridge Advanced Programme on Rethinking Development Economics – CAPORDE, University of Cambridge, July.
Oates, W. (1972), *Fiscal Federalism*, New York: Harcourt Brace Jovanovich.
Oates, W. (1999), 'An essay on fiscal federalism', *Journal of Economic Literature*, **37** (3), 1120–49.
Ottaviano, G. (2003), 'Regional policy in the global economy: insights from the new economic geography', *Regional Studies*, **37** (6–7), 665–73.
Ravallion, M. and G. Datt (2002), 'Why has economic growth been more pro-poor in some states of India than others?', *Journal of Development Economics*, **68**, 381–400.
Reserve Bank of India (2006), *Handbook of Statistics on the Indian Economy*, Mumbai, India.
Rodrik, D. and A. Subramanian (2004), 'From "Hindu growth" to productivity surge: the mystery of the Indian growth transition', IMF Working Paper 04/77.
Sachs, J.D., N. Bajpai and A. Ramiah (2002), 'Understanding regional economic growth in India', Center for International Development at Harvard University, CID Working Paper No. 88.
Sánchez-Reaza, J. and A. Rodríguez-Pose (2002), 'The impact of trade liberalization on regional disparities in Mexico', *Growth and Change*, **33** (1), 72–90.
Schleifer, A. and R. Vishny (1993), 'Corruption', *Quarterly Journal of Economics*, **108** (3), 599–617.
Singh, N. (2006), 'State finances in India: a case for systemic reform', MPRA Paper No. 1281.
Treisman, D. (2002), 'Decentralization and the quality of government', mimeo, University of California, Los Angeles.
Winkler, D. and A. Gershberg (2000), 'Education decentralization in Latin America: the effects on the quality of schooling', LCSHD Paper Series No. 59, World Bank.
World Bank (1994), *World Development Report 1994: Infrastructure for Development*, Washington, DC: World Bank.
World Bank (2004), *Investment Climate Assessment*, Washington, DC: World Bank.
World Bank (2006), *World Development Report 2006: Equity and Development*, Washington, DC: World Bank.
Zhang, X. and K.H. Zhang (2003), 'How does globalisation affect regional inequality within a developing country? Evidence from China', *Journal of Development Studies*, **39** (4), 47–67.

25 Economic decline and public intervention: do special economic zones matter?

Peter Friedrich and Chang Woon Nam

25.1 Critical assessment of development theories

Decline as development

The attempts to define development normally comprise three main features. The narrowest approach is the interpretation of development as a process of economic development like growth. Development is often associated with the achievement of economic goals. In this context it is argued that economic development is a process whereby an economy's real national income increases over a long time (Meier and Baldwin, 1957). Furthermore the concept of development is expanded under the consideration of other types of goals like sustenance, self-esteem and freedom (Goulet, 1971; Sen, 1999; World Bank, 2000). A number of non-economic indicators are also adopted when describing the state of an economy, such as the Human Development Index and the Human Poverty Index (UNDP, 2001), to name a few. As a consequence most analyses on development have a strong interdisciplinary character, and the development theory cannot solely be an economic one (Szirmai, 2005). At least a broadening of the economic base in the sense of economic goal achievement seems necessary or welcome to enhance the results with respect to other goals. Looking from the perspective of how world welfare should improve, one can also derive some similar but specific statements for regional welfare.

The theories aimed at investigating and examining development refer mostly to growth, while economic decline and those factors restricting economic development have not been examined exclusively. However, one can easily identify periods with economic and social decline in history, which include, for example, the late Roman Empire to medieval Europe (Seston, 1963). In addition, although a country continues to grow, some less competitive regions (most probably rural areas) are likely to suffer from decline that is forced and triggered by structural changes and emigration. Economic decline has also recently been recorded in several poor countries such as Cuba, North Korea and many African states, whereas some regions even in the US and Europe have experienced declining development from time to time. In this sense, the definition of development should concern decline as well.

This also has implications for the development policy-making. In some European countries and regions the debate has been going on about the suitable economic and social policy measures to prevent the decline process resulting from the population decrease, for instance. Can decline be overcome? Should policy measures be more strongly directed to decline in order to minimize economic losses caused by decline? For some theorists decline means a stagnation accompanied by a slowing growth, or an economic situation where growth stops. However, the definition should also comprise economic shrinking.

Although the development theories and concepts are mostly oriented to growth, some suggest the causes of decline and/or insights on reasons of possible decline.

The following factors of decline can be identified from the different development theories.

The neo-classical growth theories purely deal with economic factors, especially labour, capital and technical progress. Decline comes about if one of the three production factors fails to grow and cannot be compensated by the growth of the other factors, or if the growth of all three factors is negative. Government activities can also be integrated as public capital input or regional infrastructure (Timm, 1963; Andersson, 1981; Van Rompuy, 1981). Therefore, the reduced state activities or the resulted lower productivity can also lead to economic decline.

The so-called Harrod–Domar type of theories postulate that growth under price stability takes place if the natural rate of growth equals the warranted rate, and they are same as the actual growth rate. Decline in production factors may lead to a decrease of natural and actual growth rates and implies a rapid drop of marginal productivity of capital, severe budget deficits (Timm, 1963), and also reduction of public services and their productivity. If the minimum public utility level required for private sector production does not prevail, a production collapse tends to cause decline (Biermann and Friedrich, 1972). If growth factors stressed in the new growth theory – infrastructure-related increasing returns to scale, synergy effects, learning by production activities, and so on – are neglected, the decline process can also be accelerated.

The classical development theories stress the ongoing division of labour and capital accumulation which continues until a steady-state condition is achieved. If the volume of production factors shrinks because of unexpected events, climate conditions, and so on, a decline process can be initiated. The Keynesian theory suggests some frictions in the different types of markets (goods, capital, money and labour markets) which also cause decline.

Malthus explicitly explains that the mismatch between the increase in production volume and population size occurring beyond equilibrium is a cause of economic decline. Hansen (1939) argues that some demographic and economic factors like ageing, stagnating or declining population, a slower expansion of the market areas and lower demand as a consequence of higher savings would well discourage economic activities, entrepreneurship, private and public investments and innovation and technology development – all leading to economic decline. In a declining phase of a region only the drastic behavioural adjustments of economic agents and sectors appear to stimulate regional governments to react and to develop new strategies (Friedrich, 1987).

Some theories assess a fast population increase as detrimental, since this leads to a reduction in savings and consequently investments, environmental effects of overcrowding, insufficient food supply, congestion, infrastructure scarcity, and so on. In particular environmentalists stress the limits of development because of climate change, deforestation, land degradation, and so on. Therefore growth should be sustainable in the sense that it should not destroy the environmental basis and natural capital, otherwise decline follows (Pike et al., 2006).[1]

The sector structure approach to development dealing with the agricultural and urban sectors suggests the possibilities of decline if the surplus of the agricultural sector is insufficient to develop the urban sector equipped with manufacturing and service

industries. In case of non-existing agricultural surplus, inflows of goods, labour and knowledge from abroad may assist the start of growth. An unbalanced growth between both sectors might also lead to the causes of stagnation related to overpopulation, and so on. Stagnation effects may stem from private–public sector disproportions, too (Felderer, 1981).

Insufficient big investment pushes in strategic sectors, dominance of key sectors with structural weaknesses and lacking entrepreneurship may be responsible for decline as well. Those evolutionary theories explain how the life cycle of various products and sectors could well lead to overspecialization, backwardness and difficulties in adopting changes at the saturation stage (Kaniss, 1981). The way this process leads to the regional distinction of located firms, sectors, and so on is shown in the theories of regional competition (Batey and Friedrich, 2000). The theories of sector change stress the crucial role of research and development (R&D) activities for global international competition. Other sector-specific factors of decline are a less competitive, conventional educational system, less efficiency in knowledge transfers and the accumulation of qualified human capital, a slow modernization in agriculture due to weak entrepreneurship, small-sized farm entities and failures of adjustment in the coal and steel industries (in Europe), weakness of the service industry and venture capital provision, and so on.

Among the spatial change theories, the central places theories show how economic decline of central places occur if the delivery distances become larger and/or if shipping costs vary. If population size is stable and migration is possible, equilibrium is attained with the centres of large regions larger than those of small regions (Beckmann, 1981). If the size of population declines, some places lose centrality while others win (Feng and Yang, 2007). As the decline processes continue in certain regions and central places, a geographic polarization emerges (Hirschman, 1958).

If factors enforcing agglomeration become unfavourable in the course of time, the agglomerate loses its attractiveness, as agglomeration advantages from cost savings, geographic centrality, special infrastructure, highly skilled labour and entrepreneurship experiences, access to innovations, complementary industries (Beckmann, 1981) and the synergy effects among them gradually disappear, which leads to nationwide decline. The growth-poles can turn into the poles of decline if local industries get old or close down; their products become obsolete but future-oriented innovation activities are somehow neglected. With ageing complexes or clusters the synergy effects become less significant as production shrinks and the regions lose attractiveness in regional competition. The existing social, political and economic networks sometimes hinder the cluster from reacting in a timely fashion to new challenges. These weaknesses can be cushioned by foreign trade, foreign finance and migration. Foreign direct investment (FDI) may assist to close a technological gap. However, the strong involvement of foreign firms in the regional economy can hinder the endogenous development of local industries, agriculture, and so on.

The theories of public choice and bureaucracy (World Bank, 1997) indicate that decline can also occur in a democratic voting system when the median voters with particular interest in equality, leisure, low fertility, smooth life, redistribution (Neubauer, 2007) may have a decisive influence on determining public actions. In a federal system with different political parties in power at different tiers of government, the vertical as well as the horizontal negotiations regarding the necessary development policy measures against decline might be difficult (Friedrich and Feng, 2002). Frey (1968) has shown how the bottlenecks

in infrastructure or provision of goods by the private sector lead to voters' responses which have initiated a chain of government changes. In many cases the required policy decisions have been hindered and manipulated by the strong influence of some specific interest groups.

Sometimes economic decline is the consequence of government failures and inadequate laws. The mismatches in the allocation of governmental competences (for example over-centralization; see Wallerstein, 1974; Beneton, 2002), the intergovernmental fiscal relations in a country, the organization of public administrative units, the method of administrative coordination and existing bureaucratic routines, the utility functions of civil servants and managers and their goals (Niskanen, 1971; Friedrich, 1983), as well as negative consequences like corruption, political instability, and so on (Szirmai, 2005), may lead to depressive situations as well. Some authors stress the difficulties that emerge when altering an institutional path (Myrdal, 1957; North, 1993). Moreover, these factors may play even a more important role if there is an oligarchy in power, a one-party system or a dictatorship. Quite often such political orders are also based on religious or socialistic beliefs. As an exclusive theoretical model which adequately explains the decline of such systems does not exist at present (Szirmai, 2005), a series of common opinions regarding the factors of decline are offered. They include, for example: inefficient planning systems; a lack of market-oriented policy coordination and poor quality-oriented production; suppression of the wishes and needs of individuals; primary orientation of production to military purposes; no recognition of external effects; low entrepreneurship; and so on. Yet, if the entire society suffers from the serious decline process, and its transformation to the market-oriented system is hardly possible without changing the persisting power structure, such reforms could even further worsen the country's economic situation. In this context the transformation theories should ideally deliver a differentiated explanation regarding the paths into or out of decline.

Several development theories of social and economic order stress the importance of cultural factors or social capital (Westlund, 2006)[2] as the factors shaping development, while in some countries the reactions against Western penetration have led to decline (Szirmai, 2005). In Social Darwinism (Spencer, 1882) and the explanation of changes in social relations (including family structures) by Tönnies (1887), negative consequences of social transformation followed by the disintegration and erosion of norms clearly suggest the decline process initiated by some cultural factors (Roscher, 1882; Szirmai, 2005).

According to the Marxian growth economists a situation of stagnation can be overcome by a political revolution and a process of adaptation that eventually ends in a new revolution until the class warfare disappears. However this theory does not explain decline. Economic decline emerges if through the revolution the class of those who have the technical skills and the entrepreneurial knowledge is abolished and cannot be simultaneously substituted for. A series of such type of revolution further causes decline. Another type of economic social theory is that of Schumpeter in which the pioneering entrepreneurs introduce new goods, production methods, and so on, to change the economy.

The fundamental aspects of 'old' stage theories of economic development, which originated from the German historical school of political economy, are that the economic constitution of society at any epoch should be regarded as passing through a series of phases correlative with the successive stages of civilization from a hunting- and

agriculture-based society to a dual society with urban centres and growing division of labour and trade, with the phases of industrialization and establishment of commercial relations with the various, previously separated, parts of the world (Roscher, 1861; Schäffle, 1873), eventually ending up in the economic phases of mass consumption and worldwide trade (Rostow, 1960). In particular, Ritschl (1927) has built a bridge between location and regional development and development of society.

A theory of decline stages does not exist yet, because the theories mentioned above offer causes of decline but not a chain of decline stages. A few attempts to formulate the declining processes are made by historians (Spengler, 1947). These theories are to be distinguished from the conventional stage theories which are mostly concerned with economic factors and, consequently, attempt to explain economic growth.

Stages of regional development in regional decline

Economic development as well as industrial specialization patterns among a number of countries are most widely examined by modern stage theorists through adopting a simple (but practical) way of determining where the economy or the life cycle of leading industries of one country is currently positioned in the past growth-path of a more developed country. Modern stage theorists combine some of the relevant factors mentioned in other development theories (that is, theories of structural change, and development theories of foreign trade and regional competition) with the idea of stage development. 'A developing country, in an open economy context, industrialises and goes through industrial upgrading, step by step, by capitalising on the learning opportunities made available through its external relation with the more advanced world' (UNCTAD, 1995, p. 259). In other words, apart from the changes in the life cycles of dominating industries over time (for example, from a concentration on labour-intensive textile industry, steel and chemical industries to automobiles, and so on) and, consequently, in the domestic industrial structure of a country in the course of economic growth, such a development stage analysis model also provides in a regional hierarchy framework an explanation for the industrial relocation from a developed country to a less developed one through trade and foreign direct investment in response to a shift in competitiveness (Akamatsu, 1961; Kojima, 2000; Nam, 2006). Analogously this concept can also be applied to the subnational levels (regions and municipalities) that are to be found on the continuum within and between the stages.

In combination with two kinds of markets (that is, domestic and export markets) and five types of industries (that is, R&D-intensive and easily imitable high-tech industries, as well as capital-, labour- and natural resource-intensive industries), the stages of economic and industrial development can generally be divided into three phases, through which countries progress:

- stage 1: natural resource and labour driven;
- stage 2: capital and imported technology driven; and
- stage 3: R&D and innovation driven.

This definition applies to all countries which have well-functioning domestic markets that are also effectively integrated into the world market. If domestic markets hardly exist in a country because they are substituted by a centrally planned coordination of economic

units, the country is likely to be in a low stage of development by definition, although such countries may have well-developed high-tech sectors of the economic military complex. Similar to most of the development theories, a tendency to growth is assumed when moving from one stage to the following one.

According to this stage approach, each nation is on a continuum within one of these three stages, and as it moves forward, it takes on a new series of competitive tasks in the world economy and leaves less sophisticated activities to countries at the lower level of economic development. The natural resource and labour-driven stage of economic development includes countries that generate most of their gross domestic product (GDP) from processing and exporting natural resources and agricultural products. In addition, cheap, manual-skilled labour in these countries host a variety of simple mass-production assembly plants.

In the second stage, countries are more technologically advanced than countries in the first stage. Domestic and foreign investments are funnelled into plants, taking advantage of scale economies, using transferred technology from more advanced countries, and producing standardized products with mass labour inputs provided by the local population. In other words, industrial production in the stage driven by capital and imported technology is also, to a large extent, labour-intensive, and its success strongly depends on the endowment of manual and skilled workforces and their absorption capacity for foreign technology.

In the third, R&D and innovation-driven stage, firms are challenged by the increased levels of world competition to innovate new products derived from high levels of technology and know-how. Apart from the well-known impacts of modern R&D infrastructure and high-quality human capital in generating and implementing new technologies in the development of new products (Ranis, 2004), the innovative industrial firms' (institutionalized and therefore long-lasting) networks with research institutions and high-tech business service firms as well as other industrial companies in the context of a national innovation system become crucial for the country's continued economic and industrial growth in the third stage.

However, a sharp separation among the three development stages is weakening as these phases now overlap, due partly to the rapid integration of the world market and the intensive globalization of business activities of multinationals including trade, foreign direct investment and technology transfer. Moreover it is likely that the 'innovation–imitation lag' (UNCTAD, 1996, p. 80) between advanced countries (like Japan, Germany and the US) and Asian newly industrialized economies (NIEs), for example, has been significantly reduced since the mid-1980s, thanks mainly to the greater flexibility and divisibility in production technology and to a rapid accumulation of physical and human capital in NIEs that has enabled them to introduce new technologies embodied in capital goods and has accelerated the learning and catching-up process.

One critical point regarding these stage theories is that there must be indicators for markets and industries as well as values that are allocated to the individual stages, allowing the identification of the different development stage where a country or a region is positioned. In addition the formulation of indicators seems to be difficult, since they can be expressed in absolute terms or in relation to a leading country. Moreover, there is a question of how newly developing industries or technologically changing industries such as agriculture are considered within the indicator system applied. As already indicated

above, the current approach mainly concentrates on economic factors of growth and decline, encompassing a number of relevant factors examined in different types of theories. Therefore, a more systematic general theoretical concept appears to be necessary in order to explain the stage movement caused by the long-run decline, as happened in the UK between 1870 and 1950 (Hoffmann, 1955; Meier and Baldwin, 1957; Kuznets, 1966), Germany between 1900 and 1960 (Hoffmann, 1965) and China between 1400 and 1920 (Szirmai, 2005). As Britain and Germany fell back from the third stage to the second in the years mentioned above, China went back from the third stage as technologically leading in the fourteenth century, to the first stage at the end of the nineteenth century.

Moreover the new stage models should be expanded under the consideration of the different types of economic coordination systems as an additional economic performance indicator. We suggest considering the market coordination that implicitly underlies the modern stage ideas; an economy in transition from central planning to market coordination in the private sector presently existing in countries like China and Vietnam (as well as most countries in first phases of transition), and the absence of market coordination in countries like North Korea.

In general, countries in the first development stage are less developed, and therefore a wide range of policy measures could be applied to overcome such underdevelopment. Yet a selection of a suitable policy-mix appears to be necessary under the adequate consideration of their specific, weak economic structure. Those countries having reached the second stage or having fallen back to this stage primarily have to foster strategic measures to provide capital (also venture capital), to enforce technological development and to import technological knowledge. Countries falling backwards in position within the third stage should push technological development, cooperate with industrial partners in other advanced countries, develop modern service industries and concentrate more on technological strategic branches that fit into their overall economic structure.

Public interventions to overcome economic decline

In order to overcome economic decline and to achieve a higher stage of development, governments in most countries have implemented a variety of policy programmes. They comprise direct measures to influence development, such as provision of public services, public research and public production, and indirect measures of development policy concentrating on framework conditions for private and public economic units, for example competition laws.[3] With regard to market economies most economists agree on the fact that governments should always attempt to provide better legal, administrative and institutional frameworks for economic actors through indirect measures that allow for the improved compensation of market failure and the avoidance of government failure. Furthermore, thorough development, implementation and coordination as well as monitoring and evaluation of the economic policy-mix are commonly required. There is a series of economic policy measures which can be adopted to fight against economic decline directly and indirectly (in combination with demographic and social policy types) (see Table 25.1). These economic policy measures affect the economic behaviour and decision-making of firms and consumers in a dynamic process which, in turn, triggers changes in the size of individual macroeconomic aggregates and eventually stabilizes the total amount of economic output of a nation.

Table 25.1 Economic policy-mix

Macroeconomic policies	Meso-economic policies
Development measures in a market economy	
Fiscal policy	Structural policy
Monetary policy	Technology policy
Labour market policy	Regional policy
International economic policy	Environmental policy
Dual transition economy	
Fiscal policy	Structural policy
Monetary policy	Technology policy
Labour market policy	Regional policy
International economic policy	Environmental policy
Ownership policies	Ownership policies
Coordination policies	
Centrally planned economy avoiding transition	
Fiscal policy	Most notably FEZ-oriented:
Monetary policy	Fiscal policy
Labour market policy	Monetary policy
International economic policy	Labour market policy
Ownership policies	International economic policy
Coordination policies	Ownership policies
	Coordination policies
	Structural policy
	Technology policy
	Regional policy
	Environmental policy

Transition countries have to shape development measures that fit into a dual economy which exists for a certain transition period. For example, some policy differentiations appear to be desirable for these economies including: (1) a different taxation system for company owners, private domestic units and joint ventures in the form of FDIs (Friedrich, 1991); (2) a split social insurance system for employees in state-owned firms and those working in private firms or joint ventures; (3) a split ownership system; and (4) a regional split of economic orders, as seen among mainland China, its coastal provinces and Hong Kong, for example. Implications of this kind of differentiated development policy are not fully examined. However, the respective programmes consider special mixes of policies mentioned in Table 25.1.

Especially serious difficulties seem to emerge when formulating policies for those low-development-stage countries that actively pursue rapid economic development and attracting FDI, but simultaneously try to minimize changes to their social and economic command system. These countries are mostly attracted to establish regionally isolated development islands where a mix of policies according to Table 25.1 is executed, but maintaining the full control regarding the relations between activities in these geographically limited areas and those of the majority of regions of the country.

A compound measure of public development policies which is adopted in all three types of countries mentioned in Table 25.1 is the establishment of free economic zones (FEZs) that are more or less integrated into the domestic economic system, while creating and offering somewhat different conditions compared to other economic zones and regions within a country. The setting-up of FEZs can be assumed to be a direct measure, because they often have a character of a public institution or their foundation is mainly based on public initiatives. At the same time they belong to the indirect development measures, since they offer a special policy-mix integrating development measures to create an advantageous economic framework. Repeatedly, FEZs constitute a concentration of a large number of development policy measures and a big regional push of public intervention. Therefore they might be useful to demonstrate the effectiveness of public intervention to overcome decline.

25.2 Free economic zones as development strategies

Free economic zones as an instrument of economic growth and transformation

The FEZ is a territorial enclave in which foreign firms (in many cases also in cooperation with indigenous companies) benefit from generous incentives and privileges and thereby produce industrial goods mainly for export:

> There is no clear-cut definition of FEZs. [FEZs are geographically defined areas] within which certain types of economic activity take place without some of the government taxation and regulation that applies to them in the rest of the economy. [FEZs] are designated areas free of customs duties and import controls that provide an attractive environment for investment, technology, promotion of exports and employment opportunities. Free trade zones or commercial free zones . . . are warehousing areas where goods are stored and re-exported to the host country or abroad without substantial transformation. Free trade zones that include export-oriented industrial activities are named export processing zones. (Tahir, 1999, p. 3)[4]

The initial objectives of FEZs in market-oriented developing and/or emerging countries or dual transformation countries have generally been the creation of: (1) 'islands of competitiveness' in an economy which is not yet ready to submit itself fully to (market-oriented) international competition; and (2) transmitters of the advantages of the market economic system to the domestic economy to make the entire nation more prosperous and competitive (UNCTC, 1991, p. 345).[5]

The most typical and traditional ways of establishing international economic cooperation are trade and FDI, which can be located in FEZs. Investors generally tend to adopt a two-phase process when evaluating FDI locations. In the first phase they mainly examine countries and regions based on their fundamental determinants like market size, access to raw materials, availability and quality of skilled labour and infrastructure, and so on. Only those areas that satisfy these criteria go on to the next phase of evaluation, where tax rates, grants and other incentives become important. However, government can quickly and easily change the range and extent of tax incentives and other types of public promotion schemes they offer as indirect development measures bundled in an FEZ, whereas changing the other location factors mentioned above may be difficult and time-consuming, or even beyond government control (UNCTAD, 2000; Nam and Radulescu, 2004). However, to some extent direct measures such as infrastructure available in an FEZ

can at least regionally improve the first group of location conditions. Discussions concerning the efficiency of FEZs to overcome decline in market-oriented countries and in dual transformation countries concentrate on four main points: (1) policy-incentives offered in an FEZ; (2) agglomerative characteristics of an FEZ; (3) effectiveness of an FEZ as a transition instrument; and (4) practical experiences related to an FEZ.

Regarding the general effects of tax incentives and other public policy measures such as easing of foreign currency regulations, decentralization of development policy-making, and so on on firms' location in the FEZ and other types of enterprise zones, Bartik (1991) and Ge (1995) argue that there are positive relationships between the presence of such incentives and increased economic activity, which in turn helps overcome decline. It is generally argued that FEZs in these countries can lead to capital inflows, which have potential important welfare implications:

> It is believed that regulation and protection in many countries represent serious barriers to the inflow of capital. The establishment of [FEZs] which removes such barriers can induce the flow of capital [raising the labour productivity, generating linkage effects and increasing tax revenues, etc. These revenues] can be translated into gain in welfare, and, in the long run, development of the host country. (Tahir, 1999, p. 5)

In addition, as Papke (1992) suggests:

> incentive may affect factor prices, and incentives that lower the price of capital goods have both an output effect (whereby production and employment increase because costs are lowered) and a substitution effect (whereby capital is substituted for labour). If the substitution effect is stronger, a capital intensive investment could reduce employment. (Fisher and Peters, 1997, p. 125)

Unlike a larger number of references including Devereux and Chen (1995) and Schweinberger (2003) highlighting the desirability of FEZs for the stimulation of foreign capital inflows, some economists like Hamada (1974), Rodriguez (1976) and Hamilton and Svensson (1982) argue in the Heckscher–Ohlin framework that foreign investment attracted in FEZs can have an effect of lowering welfare:

> The root cause of possible immiserization effects in developing countries is that foreign investment in the [FEZs] entails movements of mobile factors between the domestic and [FEZs]. The latter may imply that the outputs of the industries which are overproduced in the original equilibrium (with economy-wide tariffs) rise even more [and consequently reduce] the value of output of the whole economy at world market prices. (Schweinberger, 2003, p. 620)

When the trade of intermediate goods is considered in the analysis, a reduction of tariffs on the import of those goods into the FEZ can again cause a decline in national income, since the distortions associated with the existing tariffs in the rest of the economy may be exacerbated (Young, 1987). In Musleh-ud (1994) where the producing sector of intermediate goods is explicitly introduced into a model of an export-processing zone, an inflow of foreign capital may increase the country's welfare if the intermediate good is non-tradable. These findings refer mainly to market integrated countries.

According to the theory of agglomeration economies, economic growth and technology development – particularly at the regional level – is influenced to overcome decline and stimulated by the economies generated by spatial proximity and associated

externalities (Glaeser et al., 1992; Mills and McDonald, 1992). By being located near various numbers and types of firms in agglomerations or FEZs, an easy and speedy business access (with low transportation costs) to other service and industrial firms (suppliers, distributors, and so on) or research institutions is guaranteed. Furthermore, in the case of expanding similar industrial branches in a given location, firms can realize economies of scale by using jointly supplied products (and raw materials) or by specializing in production. An additional benefit includes the savings resulting from intensive sharing of given major capital investment and infrastructure by a number of firms in a geographic enclave. Within an economic zone that has a concentration of rapidly growing (foreign and domestic) firms in an emerging dynamic industry and service sector, the recruitment of a specialized labour force is also convenient: modern industrial and service firms 'that are growing quickly need to be able to recruit specialised, experienced and skilled professionals who can meet specific requirements' (Mills and McDonald, 1992, p. 42). Additionally, such geographic proximity makes the inter-firm communication of new ideas, experiences and know-how among firms more efficient and innovative (the so-called Marshall–Arrow–Romer externality of knowledge spillovers between firms, according to Glaeser et al., 1992). Consequently, such advantages of agglomeration economies provided by an FEZ can have a positive effect on a local economy and stimulate efficient production and generate productivity growth, leading to higher per capita income than that in the rest of the country (Bartik, 1991). On the other hand, diseconomies of agglomeration emerge 'when concentration of population and economic activity in one place either raises the real cost of production by requiring more inputs, or reduces the real standard of living by increasing environmental, physical or social disamenities [like industrial air and water pollution, crime and traffic congestion]' (Ihlanfeldt, 1995, p. 340).

In recent years the concept of the FEZ has evolved and has been diversified. A number of export-processing zones have additionally acquired import-processing functions (see the case of the Manaus Free Zone in Brazil that now operates almost exclusively for the domestic market). Major factors that have made such trends towards import processing almost inevitably include: (1) the technical difficulty of controlling the smuggling of products and technologies from the zone into other parts of the host country; (2) the combined pressures of local consumers (who would like to have access to and can also afford the high-quality goods produced in the zone) and foreign investors (who are attracted by the potentially high profitability of sales in the local market, as is the case in China); and (3) governments' policy to encourage local linkages in exchange for access to the local market. A further important development was the establishment of domestic firms in the FEZ. In countries such as India, local participation is compulsory when a foreign firm wants to invest in the country's FEZ. This growing importance of domestic enterprises is well illustrated by the fact that over two-thirds of all enterprises located in the FEZs of developing countries are presently either fully owned indigenous firms or joint ventures between domestic companies and foreign partners. In the near future the evolution of a classical manufacturing-oriented FEZ into a modern service-oriented zone is expected. The growing service-orientation of some FEZs is, therefore, a much wider and more ambitious concept than the free ports, because it encompasses not just traditional trading and transporting activities, but also modern financial and business services such as banking, insurance and data processing. The concept of the FEZ as such a service-oriented (and

services-cum-manufacturing) zone could also encompass some tourism or educational services (UNCTC, 1991).

Major theoretical justifications for the establishment of such economic zones generally include that 'there [are] economies of scale in the development of land and in the provision of common services and utilities [and] external economies of agglomeration by having similar industries grouped together. [Furthermore] governments may wish to impose a geographical limit on the operation of some policies and to restrict certain activities to specific areas' (Wall, 1993, p. 248). For the application of the latter justification in the dual transformation countries, it is additionally suggested that, with the enclave nature of the FEZ, the process of gradually opening former command economies to the outside world can be controlled and modulated in a much more subtle and sophisticated way than through a rapid global liberalization of the total national economy (UNCTC, 1991).

Instead of being further concentrated in a well-defined territorial area, investment promotion schemes and other types of incentives provided in FEZs (like tax concessions, easing foreign currency regulations, and so on) were gradually expanded – in the course of time – to other (local or foreign-owned) enterprises, operating elsewhere in the country (see cases in Hungary). Most obviously the special economic zones (SEZs) in China have been rapidly expanded along the large coastal areas, rather than remaining as small industrial enclaves. The selection of initially four SEZs in the southern part of China in 1978/79 was mainly aimed at achieving a geographic proximity to Hong Kong, Macao and Taiwan in order to exploit fully the advantage of the highest concentration of overseas Chinese. Regarding the foreign investment activities, some significant shifts were made thereafter. These include, for example, moving away from the SEZ to a broader geographical spread leading to the subsequent expansion of SEZs along the coast, shifting concentration from real estate development (including hotels and other tourist facilities) towards industry, and turning away from joint venture-based investment to wholly owned enterprises (Wall, 1993).

However, according to past experiences worldwide, FEZs have not usually developed along the lines originally planned. Furthermore, the economic and social benefits of a zone tend to be much greater (or much smaller) than anticipated, and in most cases quite different from what had originally been planned. These facts are well indicated by the development of a number of zones into industrial monocultures, rather than into the well-balanced and highly diversified industrial parks envisaged by the planners. The phenomenon is due to a number of complex sociological and economic reasons which suggest that FEZs maintain a life of their own and an internal dynamism that one can hardly predict in the planning process. Mistakes made during the planning and design stage have also led to the failure of FEZs in many countries, which include, for example: the choice of an underdeveloped region with poor road and air communications; insufficient attention to the other basic infrastructure (such as telecommunications or electricity supply, and so on) and to the overall interregional and/or international accessibility of the region; a mismatch between skills of indigenous workforces and those required for new production activities; and so on (Chaudhuri and Adhikari, 1993). In many cases neighbouring FEZs compete against each other as potential locations for foreign firms. To a larger extent, the successful development of a zone also seems to be led by the ability and flexibility of the zone authorities to react to changing (particularly economic) circumstances, to make

the necessary mid-course corrections, to adjust the zone's institutional structure to new problems arising with zone development and, more generally, to develop an effective evaluation and problem-solving mechanism (UNCTC, 1991; Tuppen, 1993).

Microeconomic theory of FEZs

The establishment of the FEZ takes place in a framework of oligopolistic competition for development between countries under the consideration of governmental goal functions like those North Korea has had (see below). In many countries a local authority is usually established that is acting as an agent of a higher tier of government and is responsible for the start-up and the management of an FEZ. The local zone agency has to fulfil the assigned government aims and, by doing so, it can survive as an institution. Most commonly, a division of tasks takes place in the following way. Central (or provincial) government provides financial means and infrastructure access necessary for the FEZ establishment, and determines, based on its major aims, the general conditions for economic activities of firms to be located.[6] On the other hand, the FEZ management constructs the FEZ according to technical and organizational norms and builds up a negotiation network with the relevant economic, administrative and social institutions in order to create an attractive location environment for firms.

Thereafter the FEZ management attempts, in cooperation with central government institutions, to attract firms into the entity (Lindemann, 1999). These activities occur in a rather competitive environment. The market forms are mostly monopsonies where an FEZ faces a large number of competitors, although oligopsonies with a few FEZs (or other location-offering institutions) also quite often exist. Only in rare cases is there one FEZ negotiating with one firm, while it is more seldom that several firms negotiate with only one FEZ. Bilateral oligopolies appear to be hardly existent.

In order to show the ways that aims and conditions influence the outcome of settlement negotiations we first consider a simple bilateral monopoly situation. The FEZ leases or sells real estate to a locating firm. The FEZ owner (management) attempts to maximize employment (A), output (V), investments (K) and financial rewards at the same time. It offers an activity-dependent subsidy S to the firm in a form of real estate price reduction for higher economic activities (see equations 25.1 and 25.3 in Box 25.1). The private locating firm tries to maximize profits (see equation 25.4). It has a demand function (equation 25.7), a production function in which FEZ services are taken into account (W in equation 25.4) and cost of services G_w in equation (25.6) as well as a total cost function (see also equation 25.6). The W can also reflect infrastructure services provided outside the FEZ. The firm also tries to minimize costs by a given quantity of outputs and selects a combination of adequate inputs required for the production including land (B) (see also equation 25.6).

Since both utility functions commonly depend on the real estate price and output, the utility function of the FEZ and that of the settling firm can be shown as in Figure 25.1. Moreover, the figure at the top shows the possible settlement – negotiation – solutions that are Pareto-optimal for both partners (see equation 25.8 in Box 25.2). The graph at the bottom shows a Nash-solution of the settlement negotiations, which gives the utility distribution between both partners.[7] The corresponding solutions for subsidy, labour, capital, and so on are evolved (see equation 25.10 in Box 25.2).

The variables modelled in the utility function of the FEZ reflect the policy aims of government to foster employment (A) and economic activities (X), to accelerate investment

BOX 25.1 BASIC EQUATIONS OF THE FEZ MODEL

FEZ

(25.1) Value function (N_t)

$$N_t = g_{TV} \cdot V + g_{TA} \cdot A + g_{TK} \cdot K + g_{TF} \cdot F + g_{TB} \cdot S, \qquad g_{TV}, g_{TA}, g_{TK}, g_{TF}, g_{TB} > 0,$$

$$\partial N_t / \partial S = g_{TB} > 0,$$

$$dN_t/dS \Big|_{V,A,K,F_o = konstant} = -g_{TF} + g_{TB}, \qquad -g_{TF} < g_{TB} - g_{TF} < 0;$$

(25.2) Building lease

$$F = F_o - S,$$

$$\partial F/\partial S = -1 < 0;$$

(25.3) Discount lease or subsidy to promote restructuring

$$S = \eta \cdot V, \qquad \eta \geq 0.$$

Private firm

(25.4) Utility function (value, N_u)

$$N_u = (1 - t_K - t_G) \cdot (P \cdot V - C),$$

$$\text{Max} \{N_u \mid \partial F/\partial V = 0\} > 0;$$

(25.5) Production function

$$V = V_o(W) \cdot A^{\alpha} \cdot K^{\beta} \cdot B^{\gamma}, \qquad \alpha, \beta, \gamma > 0, \quad \alpha + \beta + \gamma = 1,$$

$$B = B_o + B_t,$$

$$\partial V_o(W)/\partial W > 0;$$

(25.6) Costs function

$$C = G_W + r_A \cdot A + r_K \cdot K + (r_B + t_B) \cdot B + F_o - \eta \cdot V - F_B = V/m + G_W + F - F_B,$$

$$F_B = r_B \cdot B_o,$$

$$r_A \cdot A/\alpha = r_K \cdot K/\beta = (r_B + t_B) \cdot B/\gamma = V/m, \quad \text{Expansion path}$$

$$m = V_o(W) \cdot [\alpha/r_A]^{\alpha} \cdot [\beta/r_K]^{\beta} \cdot [\gamma/(r_B + t_B)]^{\gamma},$$

$$A = V/m_A, \quad K = V/m_K, \quad B = V/m_B,$$

$$m_A = dV/dA = m \cdot [r_A/\alpha],$$

$$m_K = dV/dK = m \cdot [r_K/\beta],$$

$$m_B = dV/dB = m \cdot [(r_B + t_B)/\gamma],$$

$$1/m = \partial C/\partial V, \quad \text{if} \quad \partial F/\partial V = 0;$$

(25.7) Demand function

$$P = P_o - P_t \cdot V.$$

Source: Feng and Friedrich (1995).

(K) and also to promote technological development at most favourable fiscal conditions provided in the FEZ. The solution formulas are demonstrated in Box 25.2 for the variables relevant for the agency such as utility, activity level, subsidization, real estate price, employment and investment. Within this model the minimum conditions of utilities of the negotiating parties, namely N_{uo} (firm) and N_{to} (FEZ), are determined. They occur in the resulting formulas for their solutions as well (see equations 25.9 and 25.10). Graphically they occur in Figure 25.2 as lines N_{uo} and N_{to} suggest.

Such a model approach opens possibilities to examine the potential competition against other FEZs. A settlement firm negotiates with the FEZ_1 in country 1 and the competing FEZ_2 in country 2. The settlement firm carries out the first round of negotiations with the FEZ_1 of which result comes to a solution shown in Figure 25.2. In the next round it deals with the FEZ_2 asking for a minimum utility as high as the utility level achieved in the solution with the FEZ_1. A Nash solution shows the utility level of the firm reached with the FEZ_2. Again this utility level is used as the minimum claim in the next round of negotiation with the FEZ_1 and so forth, until one of the competitors cannot offer further utility improvements for the firm. The FEZ finally wins, which provides together with the settlement firm the larger common utility space for mutual solutions.

Based on such a model the consequences of declining development on the outcome of economic promotion policy in a form of establishing the FEZ can also be investigated. If population declines, the domestic market is in trouble. The demand curve of the entrepreneur for the products provided in the FEZ country shifts 'inwards', which means that P_o gets smaller. Factor lines with shrinking purchasing power due to the income decrease (as a consequence of economic declines) as well as the shortage of foreign currency restricting demand for foreign intermediate goods, and so on, lead to the smaller solution values (see equations 25.9 and 25.10 in Box 25.2). There is an inward shift of the solution space (see equation 25.8). Accordingly the business promotion turns out to be less effective than hoped for by FEZ_1. If the demand reduction is extremely high, it tends to

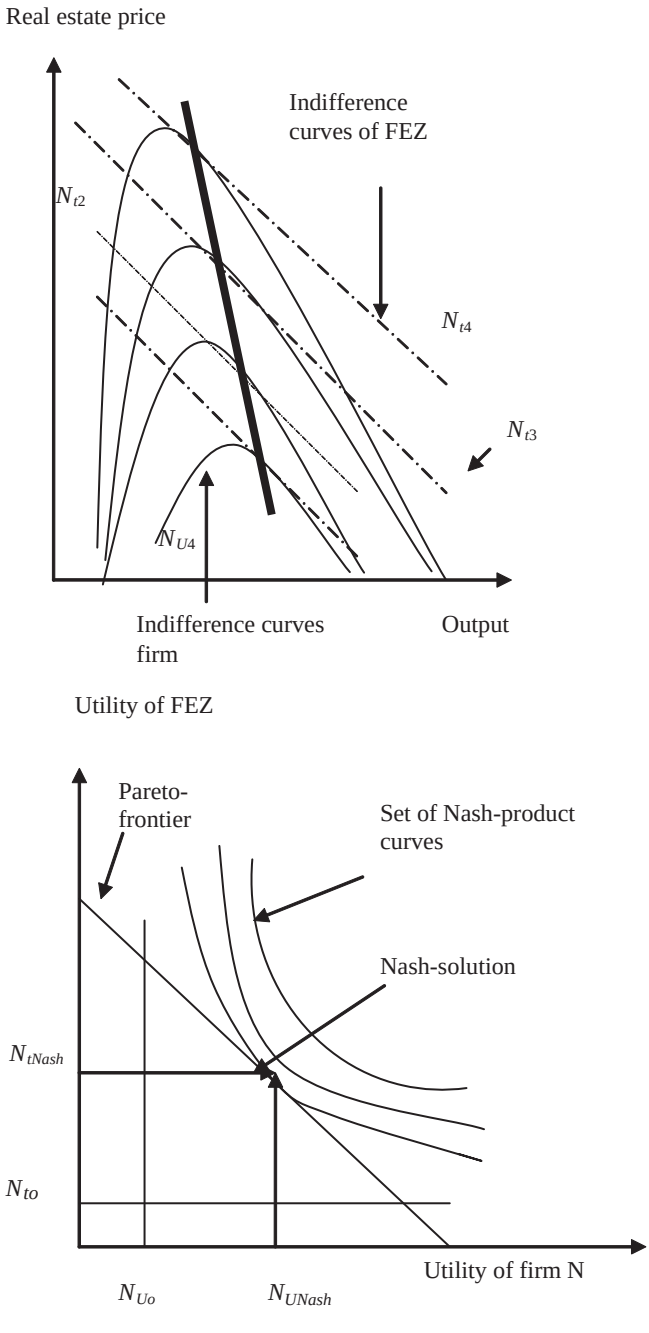


Figure 25.1 Pareto-optimal contracts between FEZ agency and settlement firm and negotiation solution

BOX 25.2 NASH SOLUTIONS FOR VARIABLES AND SETTLEMENT CONTRACT STIPULATIONS

(25.8) Pareto solution

$$V = \frac{1}{2 \cdot P_1} \cdot \left[P_o + \frac{1}{g_{TF}} \cdot \left(g_{TV} + \frac{g_{TA}}{m_A} + \frac{g_{TK}}{m_K} + g_{TB} \cdot \eta \right) - \frac{1}{m} \right],$$

$$dV/d\,\eta = g_{TB}/(2 \cdot P_1 \cdot g_{TF}) > 0;$$

$$\eta = -2 \cdot P_1 \cdot (g_{TF}/g_{TB}) \cdot V + (g_{TV} + g_{TA}/m_A + g_{TK}/m_K)/(g_{TF} - g_{TB}),$$

$$\text{with} \quad (g_{TV} + g_{TA}/m_A + g_{TK}/m_K)/(g_{TF} - g_{TB}) - 1/m < 0;$$

$$\frac{N_t}{g_{TF}} + \frac{N_u}{1 - t_K - t_G} = \frac{1}{16 \cdot P_1} \cdot \left[P_o + \frac{1}{g_{TF} - g_{TB}} \cdot \left(g_{TV} + \frac{g_{TA}}{m_A} + \frac{g_{TK}}{m_K} \right) - \frac{1}{m} \right]^2 - G_W + F_B .$$

(25.9) Nash solution concerning utility split

$$N_{tNash} = \frac{g_{TF}}{2} \cdot \left\{ \frac{N_{to}}{g_{TF}} - \frac{N_{uo}}{1 - t_K - t_G} + F_B - G_W + \frac{1}{16 \cdot P_1} \cdot \left[P_o + \frac{1}{g_{TF} - g_{TB}} \cdot \left(g_{TV} + \frac{g_{TA}}{m_A} + \frac{g_{TK}}{m_K} \right) - \frac{1}{m} \right]^2 \right\},$$

$$N_{uNash} = \frac{1 - t_K - t_G}{2} \cdot \left\{ \frac{N_{uo}}{1 - t_K - t_G} - \frac{N_{to}}{g_{TF}} + F_B - G_W + \frac{1}{16 \cdot P_1} \cdot \left[P_o + \frac{1}{g_{TF} - g_{TB}} \cdot \left(g_{TV} + \frac{g_{TA}}{m_A} + \frac{g_{TK}}{m_K} \right) - \frac{1}{m} \right]^2 \right\}.$$

(25.10) Nash solution concerning contract stipulations

$$V_{Nash} = \frac{1}{4 \cdot P_1} \cdot \left[P_o + \frac{1}{g_{TF} - g_{TB}} \cdot \left(g_{TV} + \frac{g_{TA}}{m_A} + \frac{g_{TK}}{m_K} \right) - \frac{1}{m} \right],$$

$$\eta_{Nash} = -\frac{g_{TF}}{2 \cdot g_{TB}} \cdot \left(P_o - \frac{1}{m}\right) + \frac{1}{2} \cdot \left(\frac{1}{g_{TF} - g_{TB}} - \frac{1}{g_{TB}}\right) \cdot \left(g_{TV} + \frac{g_{TA}}{m_A} + \frac{g_{TK}}{m_K}\right),$$

$$S_{Nash} = \eta_{Nash} \cdot V_{Nash} = \frac{1}{8 \cdot P_1 \cdot (g_{TF} - g_{TB})^2} \cdot \left(g_{TV} + \frac{g_{TA}}{m_A} + \frac{g_{TK}}{m_K}\right)^2 \cdot \left(2 - \frac{g_{TF}}{g_{TB}}\right) - \frac{1}{8 \cdot P_1 \cdot g_{TB}} \cdot \left\{\left(P_o - \frac{1}{m}\right) \cdot \left[g_{TF} \cdot \left(P_o - \frac{1}{m}\right) + \left(g_{TV} + \frac{g_{TA}}{m_A} + \frac{g_{TK}}{m_K}\right) \cdot \frac{2 \cdot g_{TB}}{g_{TF} - g_{TB}}\right]\right\},$$

$$F_{oNash} = \frac{1}{2} \cdot \left[\frac{N_{to}}{g_{TF}} - \frac{N_{uo}}{1 - t_K - t_G} + F_B - G_W\right] + \frac{5 \cdot g_{TB} - 4 \cdot g_{TF}}{32 \cdot P_1 \cdot g_{TB}} \cdot \left[P_o - \frac{1}{m} + \frac{1}{g_{TF} - g_{TB}} \cdot \left(g_{TV} + \frac{g_{TA}}{m_A} + \frac{g_{TK}}{m_K}\right)\right],$$

$$F_{Nash} = \frac{1}{2} \cdot \left[\frac{N_{to}}{g_{TF}} - \frac{N_{uo}}{1 - t_K - t_G} + F_B - G_W\right] + \left[P_o - \frac{1}{m} + \frac{1}{g_{TF} - g_{TB}} \cdot \left(g_{TV} + \frac{g_{TA}}{m_A} + \frac{g_{TK}}{m_K}\right)\right] \cdot \frac{5 \cdot g_{TB} - 4 \cdot g_{TF}}{32 \cdot P_1 \cdot g_{TB}} - \frac{1}{8 \cdot P_1 \cdot (g_{TF} - g_{TB})^2} \cdot \left(g_{TV} + \frac{g_{TA}}{m_A} + \frac{g_{TK}}{m_K}\right)^2 \cdot \left(2 - \frac{g_{TF}}{g_{TB}}\right) + \frac{1}{8 \cdot P_1 \cdot g_{TB}} \cdot \left\{\left(P_o - \frac{1}{m}\right) \cdot \left[g_{TF} \cdot \left(P_o - \frac{1}{m}\right) + \left(g_{TV} + \frac{g_{TA}}{m_A} + \frac{g_{TK}}{m_K}\right) \cdot \frac{2 \cdot g_{TB}}{g_{TF} - g_{TB}}\right]\right\},$$

$$A_{Nash} = \frac{V_{Nash}}{m_A} = \frac{1}{4 \cdot P_1 \cdot m_A} \cdot \left[P_o + \frac{1}{g_{TF} - g_{TB}} \cdot \left(g_{TV} + \frac{g_{TA}}{m_A} + \frac{g_{TK}}{m_K}\right) - \frac{1}{m}\right],$$

$$K_{Nash} = \frac{V_{Nash}}{m_K} = \frac{1}{4 \cdot P_1 \cdot m_K} \cdot \left[P_o + \frac{1}{g_{TF} - g_{TB}} \cdot \left(g_{TV} + \frac{g_{TA}}{m_A} + \frac{g_{TK}}{m_K}\right) - \frac{1}{m}\right].$$

Note: Utility of FEZ: N_{tNash}, utility of firm: N_{uNash}, output: V_{Nash}, subsidy: S_{Nash}, real estate price without subsidy: F_{0Nash}, real estate price: F_{Nash}, labour: A_{Nash}, capital: K_{Nash}.

Source: Feng and Friedrich (1995).

happen that the minimum conditions cannot be met for one (FEZ_1) or both partners. This fact would, in turn, allow no settlement to take place.

If the lagging development increases the costs of pre-services and infrastructure we would experience an analogous result as G_w also increases. The solution space shrinks (see equation 25.8). The rewards from the sale and the utility of the involved partners at the FEZ_1 shrink as well (see equation 25.9 and 25.10 in Box 25.2). Again the opportunities to overcome negative developments through the settlements in the FEZ_1 would become considerably less favourable.

The reduced solution space can prevent a contract and/or lead to a defeat for the FEZ_1 in competition. Another condition which has a negative implication is that the real estate in current use goes up. This is reflected in a smaller m, thus reducing the space of possible solutions again (see equation 25.8). The utilities decline according to the Nash solution (equation 25.9), the fiscal financial reward F from the real estate sale, but also the output, and the factor inputs (see equations 25.10) shrink correspondingly. In this case the promotion policy implemented by the FEZ authority would remain less effective and the losses in regional settlement competition are more likely to occur.

Factor prices for labour and capital could increase, because of the lack of skilled workers or specialists who have to be recruited from abroad, and because of scarce infrastructure such as housing, social life, and so on. These tendencies are demonstrated by a change of m_A and m_K. As the labour and capital costs grow, the solution space and the utilities shrink and the outputs, employment and investments decrease consequently. Again the FEZ_1 faces disadvantages in regional competition against other neighbouring FEZs.

Tax concessions like tax holidays and tax preferences increase the utility of the settlement firm (see equations 25.8 and 25.9) and thus the solution space, and improve the competitive position of the FEZ_1. However, it does not influence the solution value of other relevant variables as strongly as is often assumed (see equations 25.9 and also Nam and Radulescu, 2004). Unstable legal system caused by a large number of changes in law, administrative weaknesses in the planning and realization process, bureaucracy and hidden corruption, and so on, can also rapidly increase the costs which can be reflected in higher factor prices, insurance costs and different demand situations. As far as these circumstances reflect declining development, the FEZ-oriented policy appears to become less effective.

In order to overcome economic problems governments can also react by adopting various strategies such as: further reduction of tax burdens; lowering tariffs; provision of more favourable infrastructure aimed at indirectly reducing factor costs for production; creation of a better investment environment including quality improvement in amenities, establishment of towns for foreigners, and so on. If the goal function of the acting FEZ as agent is less directed to financial revenues but more to development, the term g_{FT} gets reduced, while g_{TB} gets bigger. This fact implies that: (1) the indifference curve of the FEZ gets steeper; (2) the Pareto-optimal points more often correspond with higher outputs; and finally (3) the utility possibility curves shift outwards. In addition, there is a tendency to lower prices for real estate, and so on. Yet these policies appear to be rather costly and in the course of economic shrinking less feasible, since insufficient capital is available to overcome the disadvantages for the settlement firms. In this case an attempt of introducing FEZ policy would again remain less successful than originally anticipated.

The example of Najin–Sonbong FEZ in North Korea

North Korea shows most of the decline factors we identified from the development theories mentioned above. There are presently serious shortages of production factors and low productivity levels, no significant growth, low synergy effects caused by poor infrastructure endowment, environmental problems, a decrease in population, unbalanced sector developments and industry–service mismatches, shortage of agricultural goods, low adaptation of technical progress, planned development of agglomerations, cluster and growth poles, inefficient economic planning systems, suppression of wishes and needs of individuals, primary production orientation for military purposes, low entrepreneurship, and an unstable political situation. Obviously North Korea suffers from serious decline. According to the modern stage theory of development North Korea would be classified as a low-stage socialist country.

Under these conditions it seems rather doubtful whether a FEZ could help North Korea to overcome decline. Our microeconomic theory of the FEZ would lead to the following threefold results:

- The decline factors summarized in Table 25.2 with respect to declining economic conditions for settlement firms, deteriorating infrastructure, and difficult production conditions are all effectively leading to rather limited negotiation possibilities of FEZs and interested settlement firms (see also Figures 25.1 and 25.2).
- Under these circumstances a possible North Korean FEZ would not be in a favourable position to attract FDIs and to win the firm settlement competition against FEZs in other countries (see Figure 25.2).
- North Korea has to offer many additional advantages to attract firms, concerning taxation, customs, land use prices, trade possibilities, currency exchange conditions, profit transfer, and so on.

These results coincide with the empirical findings concerning the management and success of North Korean FEZs. In order to overcome economic decline the North Korean government altered at the beginning of the 1990s its negative attitude on establishing the free economic and trade zone, for political and economic reasons. These include: (1) the limited economic development caused by the centrally defined ineffective economic strategies and the existing inefficient production system; (2) the shortage of foreign currency; (3) the success of China's gradual 'open-door' transformation policy; and (4) the disruption of trade with the former Soviet Union since the end of the 1980s.

In this context, the North Korean government declared in 1991 the two cities Najin and Sonbong as the first FEZ in North Korea, which would develop – within 20 years from 1993 to 2010 as originally estimated – into not only a major logistics centre but also a modern export-oriented industrial region with additional tourist attractions.[8] The case study on Najin–Sonbong FEZ in North Korea delivers a wide range of crucial economic determinants and political reasons why according to our FEZ theory such a development intervention can fail. The development potentials of this FEZ have been badly damaged and profit-making has been difficult for foreign investors[9] there due to: (1) the long-lasting unfavourable national economic development and limited market potential; (2) the region's poor infrastructure endowment and geographic remoteness; (3) its fierce competition against those directly neighbouring Chinese and Russian FEZs along the Tumen

Table 25.2 Negotiation results under declining development

Change in decline	N_t	N_u	V	A	K	F	S	g
Declining economic conditions for settlement firm								
P_o ↓ demand	↓	↓	↓	↓	↓	↑	↑	↑
G_w ↑ costs of services	↓	↓	↓	0	0	↓	0	0
r_A ↑ labour	↓	↓	↓	↓	↑	↓	↑	↑
r_K ↑ capital	↓	↓	↓	↑	↑	↓	↑	↑↑
r_B ↑ real estate current costs	↓	↓	↓	↑	↑	↑	↑	↑
Higher ↑ firm minimum utility	↓	↑	0	0	0	↓	0	0
Deteriorating infrastructure								
Larger ↑ distance to suppliers	↓	↓	↓	↑	↑	↓	↑	↑
Larger ↑ distance to sales area	↓	↓	↓	↓	↓	↓↑	↑	↑
Higher ↑ transport costs	↓	↓	↓	↓	↓	↓	↑	↑
More difficult production conditions								
Technological parameter ↓	↓	↓	↓	↑	↑	↓	↑	↑
Exponent of labour ↓ α β	↓	↓	↓	↓	↑	↓	↑	↑
Exponent of capital ↓ β	↓	↓	↓	↑	↑	↓	↑	↑
Exponent of land ↓ γ	↓	↓	↓	↑	↑	↓	↑	↑
Reactions of FEZ								
Lowering of tax rates								
Corporation tax ↓ t_K	↑	↑	0	0	0	↑	0	0
Other profit tax ↓ t_G	↑	↑	0	0	0	↑	0	0
Real estate tax ↓ t_B	↑	↑	0	0	0	↑	0	0
Changing FEZ minimum utility ↓ N_{to}	↓	↑	0	0	0	↓	0	0
Higher ↑ output preference g_{TV}	↑	↑	↑	↑	↑	↑	↑	↑
Higher ↑ preferences for jobs gT_A	↑	↑	↑	↑	↑	↓	↑	↑
Higher ↑ preference for investment g_{TK}	↑	↑	↑	↑	↑	↓	↑	↑
Less ↓ preference for receipts g_{TF}	↓	↑	↑	↑	↑	↓	↑	↑
Higher ↑ preference to output subsidy g_{TB}	↑	↑	↑	↑	↑	↓	↑	↑

River as well as other domestic special districts (Sinuiju, Gaesong and Pyongyang–Nampo areas); (4) higher labour cost compared to that of Vietnam and interference in enterprise (including human) management by Communist Party officials; (5) hyper-inflation accompanied by inconvertibility of North Korean currency; (6) limited decentralization of economic policy-making and the fear of rapid market liberalization of the ruling regime; as well as (7) the political instability and hostility towards South Korea that have impeded the inter-Korean economic relationship (see also Nam and Radulescu, 2004).

The original schedule for the establishment of the Najin–Sonbong FEZ,[10] which was modified in 1994, consisted of three stages. In the period of 1993–95 the endowment of infrastructure would be rapidly improved as an urgent necessity for developing the FEZ as a future intermediate cargo zone. In addition, North Korea planned to establish nine industrial parks for firms and foreign investors in the FEZ, in which particularly firms in light industry (textile, toys, and so on), electronics (like TV, micro-chips, and so on) and auto parts would be located. It was, however, clear that technology and know-how, and capital and labour necessary for those industrial activities, would all have to be imported

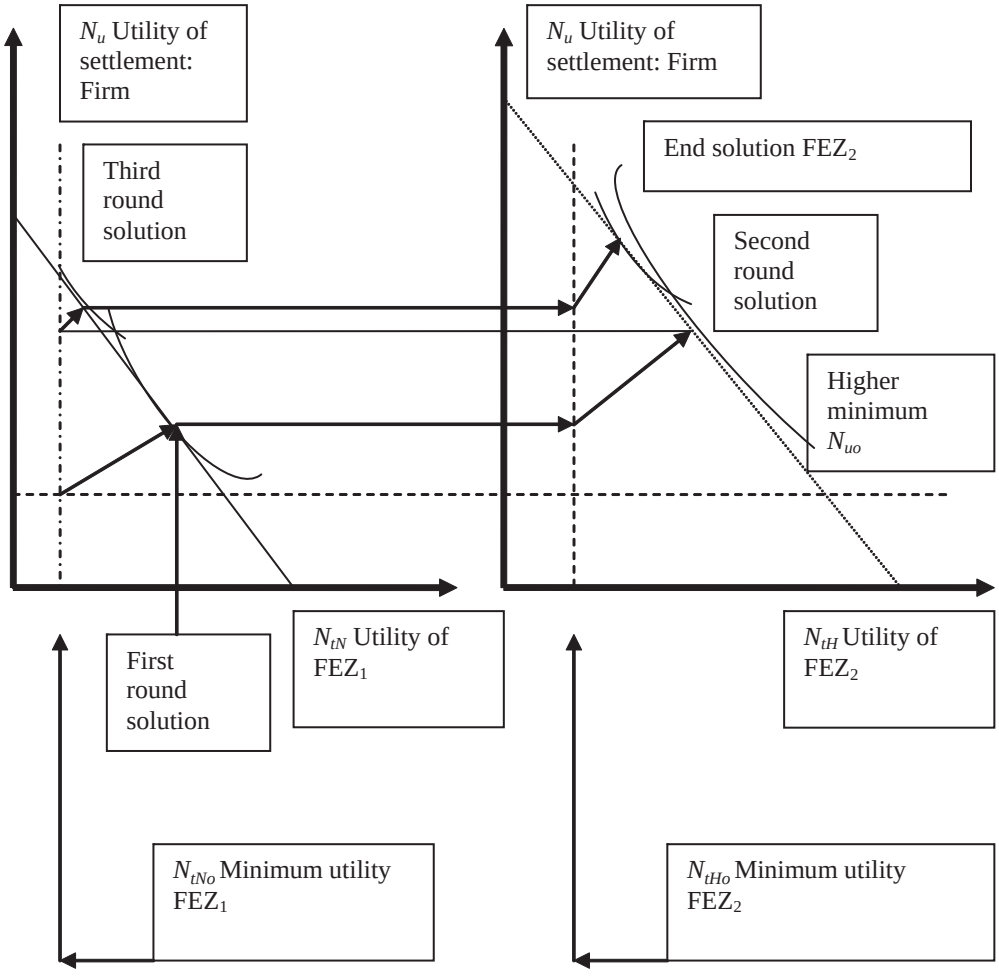


Figure 25.2 Settlement competition between FEZs

either from other parts of North Korea or from abroad, as the existing industries in the Najin–Sonbong area are poorly developed. Furthermore, the attempts of around 20 indigenous manufacturing firms (active in the fields of foods, textiles, chemical goods, and so on) to realize foreign involvement in their capacity expansion and modernization were in vain, mainly because in entering the operation phase further immense efforts were concentrated on attracting the establishment of foreign industrial firms in Najin–Sonbong and on shaping this area as a transit and export-processing zone. In the final phase from 2001 to 2010, Najin–Sonbong would be expected to grow as an international (financial and logistic) service-oriented modern manufacturing centre in north-east Asia.

In favour of Najin–Sobong in North Korea the government offered quite attractive tax investment incentives and other regulations to foreign investors in Najin–Sonbong, in many ways being more favourable than those available in China. The North Korean Free Economic Trade Zone Law allows[11] activities of the entrepreneur aimed at profit maxi-

mization and it offers: (1) a guarantee of freedom of investment and in carrying out profit-making business activities within the FEZ; (2) freedom of price determination in the context of the market mechanism; (3) the requirement that economic activity be export-oriented; (4) permission to possess 100 per cent of the capital invested and the elimination of any possibility of nationalization of foreign property; (5) the right of free choice of investment form between individual firms, joint venture and joint management; (6) the right to use land for 50 years, which can also be transferred and traded, and the possibility of extending the period of land usage; (7) permission to bring raw materials, semi-finished goods and parts for assembly into the FEZ duty-free, when the final goods will be exported; (8) permission for the tax-free overseas repatriation of profits and other legitimate revenues as well as of assets invested and the proceeds from their sale; and (9) a non-visa system permitting foreigners free access to the FEZ and free entry into the harbour for foreign ships (see also Park, 1996). While foreign enterprises are required to pay corporate income tax, the rate is set at 14 per cent within the FEZ, and 10 per cent for certain preferred high-technology activities, compared to 25 per cent and 20 per cent respectively in other parts of North Korea. Further, such enterprises (apart from those in the service sector) will be exempted from enterprise income tax for three years after the first profit-making year, provided they are operated for a period of at least ten years. A taxation reduction of up to 50 per cent (which may be discretionary) is also applicable. Other taxation incentives are offered with regard to funds reinvested or used for the provision of infrastructure. Such tax concessions are not allowed in other parts of North Korea.

Although there have been some improvements in the physical infrastructure in the Najin–Sonbong FEZ, its serious bottleneck, the interregional and international transport network, has been reduced slowly. The few projects completed include: (1) expansion of the Najin port cargo facilities; (2) construction of an international hotel in Najin; (3) electrification of the Hoeryong–Namyang–Haksong railroad; (4) expansion of the Najin–Wonjong (51 km) and the Chonjin–Hoeryong (100 km) highways and of a telecommunication centre in Najin; and (5) opening of direct container shipping services between Yanbian (China) and Pusan (South Korea) via Najin port (Kim, 1996). The major reason why North Korea was not able to improve infrastructure endowment significantly in the Najin–Sonbong area corresponding to the original schedule is that the government has fallen far short of meeting the huge capital requirements.[12] For this reason, North Korea supported the creation of the Tumen Trust Fund by the United Nations Development Programme (UNDP) and has been seeking to become a member of the Asian Development Bank (ADB), which would eventually stimulate Japanese and other international agencies in lending the necessary funds for infrastructure development in the Najin–Sonbong FEZ. However, such efforts have remained in vain.

Apart from the poor endowment of infrastructure and its remoteness, there are several economic reasons why investments of foreign firms were not strongly stimulated in the Najin–Sonbong FEZ. The Chinese rival FEZ Hunchun gained from the recent expansion and modernization of the directly neighbouring harbour in Zarubino (Russia), from which a direct railway connection was recently made to Hunchun (China), has not only endangered the role of Najin–Sonbong as the export gate for Chinese products to the world but also further reduced the region's attractiveness as the production site of international firms. Moreover, there was fierce competition among the neighbouring FEZs like Hunchun and Posyet as well as Vladivostok as potential locations for foreign firms,

whereas the efficient coordination of various interests among the neighbouring FEZs in three different countries has been difficult in the context of the international Tumen River Area Development Project. For example, according to the report of the UNDP and the UNIDO (United Nations Industrial Development Organization), up until the first half of 1995, the total foreign investment which was actually made in the Tumen River Economic Development Area region amounted to US$220 million, breaking down to US$140 million in China, US$60 million in Russia and US$20 million in North Korea. These statistics clearly show that foreign firms have strongly preferred the Tumen River regions in the former two countries. To be sure, such FDIs particularly in the fields of manufacturing processing have been not only export-oriented but also significantly affected by the size of the potential domestic markets.

Moreover, the government of North Korean develops competitors in firm settlement for Najin–Sonbong as well. Parallel to the adjustment of the Najin–Sonbong FEZ plan, North Korea has developed the city regions in the western part of the country such as Pyongyang, Nampo and Haeju as further special economic areas for attracting direct investments from South Korea, Japan, the US and Hong Kong, especially in the fields of textiles, electronics and food processing. Compared to the case in Najin–Sonbong, South Korean direct investment has been more strongly concentrated on the Pyongyang–Nampo area, which is not only more centrally located but also better endowed with infrastructure (see also Hughes, 2000). In 2002 Sinuiju, located at the north-east edge of the country bordering to China, and Gaesong, only 78 km from Seoul, were designated as additional SEZs: the latter would be particularly developed in cooperation with South Korea (Lim and Chung, 2004).

A further major difficulty involves the quality and recruitment of labour. Although enterprises in Najin–Sonbong have the right to hire and fire, an agency of the North Korean government has been responsible for providing the labour. This enables the state to keep some proportion of the wages paid to labourers. It also prevents enterprises from recruiting experienced labour from the rest of the country, or attracting individual workforces by offering suitable terms and conditions. In particular the average monthly wage of an ordinary worker is estimated to be about US$80–100 in North Korea compared to the Vietnamese level of US$50–60 in 2003 (Jung et al., 2004). Furthermore, due to the low labour productivity, the transfer of advanced technology and (international) management skills that North Korea expected to achieve through the Najin–Sonbong area were also difficult to realize in the short-term.

The North Korean currency is not freely convertible, with the won pegged to the US dollar at an arbitrary rate (US$1 = around 150 won in 2003 compared to 2.2 won in 2001). Yet the correct exchange rate of US$1 amounted to around 900 won in 2003 (Lim and Chung, 2004). Foreign currency can be brought into the country in any amount, but transactions employing the currency must all be conducted in foreign exchange coupons. These currency arrangements in combination with the underdeveloped banking system and a high annual inflation rate of 700 per cent present serious difficulties for business in North Korea. For example, foreign firms will bring capital into the country to start up the envisaged joint venture, at the present exchange rate; however, the amounts required for these reasons may be excessive.

Just as overseas Chinese have played a decisive role in the economic development in the SEZ, it appears that North Korea has strongly reckoned with large-scale investment pro-

jects in the Najin–Sonbong area initiated by South Korean firms and overseas Koreans. However, political insecurity is still the major obstacle for those South Korean firms that would like to do business in North Korea (Jung et al., 2004). Since the significant progress in attracting FDI in the first North Korean FEZ appeared to be problematic, the communist regime revised the ambitious original plan in 1994. The first stage, meant to be completed by 1995, was extended to the year 2000 to encompass the enhancement of infrastructure to foster the area as an intermediate cargo zone and an international tourist attraction. The second stage of the modified plan is to make the Najin–Sonbong area a second 'Singapore' by 2010, developing the FEZ primarily as a logistics centre, a tourist area and an export-processing base. The plans to develop it as an industrial region by establishing nine industrial complexes have been set back (Kim, 1996). In spite of the intensive efforts recently made in attracting foreign capital to the Najin–Sonbong area and the declaration of intention to create additional FEZs, many economists and politicians in South Korea and abroad interpret the setting-up of the Najin–Sonbong and other planned FEZs as a strategy of safeguarding the existing communist ruling system than the pursuit of the Chinese-style economic transformation (Jung, Kim and Kobayashi, 2004). As a consequence, it is quite often suggested that the regime in North Korea has little interest in integrating the FEZ into the rest of the economy and instead prefers the Najin–Sonbong area to remain an isolated enclave (Cotton, 1994).[13] Furthermore, it is doubtful whether the North Korean bureaucrats accustomed to a centralized administrative system would accept the decentralization of authority and the regionalization of economic policy-making that would follow the introduction of a market system as in China. Given the highly controlled political (and also economic) nature of North Korean society, it would appear to be a difficult task to make the Najin–Sonbong FEZ work efficiently and, thus, the opening-up of the North Korean economy will remain limited.

Here the FEZ is not an adequate policy option to overcome decline, not only because of the difficulties of integrating the FEZ into the planned economic system, and the fact that those advantages offered to investors do not compensate the business disadvantages, but also due to the weak financial and administrative as well as infrastructure potential of the non-transformation country of North Korea. It also shows that economic decline stemming from the general political and cultural conditions cannot largely be rectified by creating an isolated region of concentrated development policy and intervention. Yet, some positive effects can be expected if the advantages offered in the FEZ are overwhelmingly high, as our model suggests.

25.3 FEZ as a policy instrument against economic decline

Our investigation demonstrates among others:

- In the case of implementing an FEZ a concentrated programme of development measures can be launched.
- As FEZs attract internationally active entrepreneurs and investors, they have to reflect well the wishes of entrepreneurs concerning profitability, freedom and safety. This appears to be better provided and accomplished in the market-oriented countries and the dual transition countries. In those non-transition countries this kind of anti-decline policy is likely to be unsuccessful.

- An FEZ can at best help overcome decline that has economic reasons such as lack of capital and technology, persisting unemployment, lack of demand and, partly, lack of entrepreneurial initiative and/or regional competitiveness. The FEZ can also influence regional and national economic process through spillover effects, agglomeration externalities, and so on, also in terms of industrial clustering, education and the formation of human and social capital.
- The non-economic reasons for decline such as population shrinkage, political, sociological and religious factors, environmental damage and geographic remoteness can be influenced by the establishment of FEZs in a limited way.
- The need of FEZs and their success should be judged under adequate consideration of their competitive environment. In particular their impacts on domestic markets have to be thoroughly analysed. Otherwise they may not correctly pursue the initial aim of policy intervention, namely to combat decline, but may even accelerate decline when firms located in an FEZ are more competitive on domestic markets than indigenous firms in the rest of the country.
- An FEZ enhances the participation of foreign economic actors in the domestic economy, who may be less interested in whether the location country improves its position in the development stages or not.
- The type of activities performed in an FEZ are often crucial for diminishing domestic decline. In general, those complementary activities supporting domestic economic performance, or at least those which cause an influx of foreign currency, are welcome.
- In order to obtain a remarkable effect on turning decline into growth, the size of FEZs must be large.
- For identifying the ways in which the FEZs affect economies at different stages of development of growth or decline, an improved stage theory appears to be necessary to show the processes of development, and not merely a characterization of development stages based on some economic indicators. The existing development theories are less helpful to solve this task.

Notes

1. In a broader sense there must be sufficient resources and consumption goods to enable developments of new techniques and goods to achieve sustainability.
2. These include, for example, Protestant ethics, Confucianism, civic culture, caste systems in India, radical suppressing socialism and their influence on work ethos, entrepreneurship, self-responsibility, trust, division of labour, integration of women in production, subsistence agriculture, and so on.
3. In the academic and political discussions there are a large number of diverse pro- and contra-arguments about the effectiveness and desirability of such economic policy measures, which are also represented by the different groups of economic thought ranging from the classical, Keynesian and monetarist to their recent developments like post-Keynesian and neoclassical schools, to name a few.
4. In general FEZs differ from the industrial, technology and science parks in the sense that these institutions normally act under the prevailing regulations existing in a country, while FEZs benefit from the special preferences which are not provided to domestic economic units outside the FEZ.
5. It is well acknowledged that the introduction of market mechanisms and the modernization of economic structure in developing and transformation countries can be more rapidly and efficiently carried out through comprehensive integration into the international economic and business system.
6. These conditions refer to freedom of trade, acceptance of joint ventures, considering the role of foreign firms located in the FEZs in the economic planning or market system, the access of such firms to national banking and crediting, tariffs, labour conditions and staffing policies, taxation, profit transfers, exports and imports, environmental regulations, and so on.
7. The Nash solution results from maximising the Nash product under condition of Pareto-optimality (see equation 25.8). The Nash product is $(N_t\text{-}N_{to})^*(N_u\text{-}N_{uo})$.

8. Therefore we have a FEZ_1 'Najin–Sobong' that in cooperation with the central government shows a similar goal function as in the model formulated above.
9. All this leads to unfavourable conditions for the firms reflected in our settlement competition model.
10. The Najin–Sonbong region is located in the eastern part of the country and is 748 km from the capital Pyongyang.
11. The legislation relating to the FEZ distinguishes between equity and contractual joint ventures (that is, ventures managed jointly or exclusively by the Korean partner) and foreign enterprises. Foreign enterprises may only operate within the FEZ, whereas the other categories may operate in any approved venue on North Korean territory. While the regulations pertaining to the operation of equity and contractual ventures are somewhat general in nature, those regulating foreign enterprises contain a detailed account of the concessions available for such activities.
12. '[Financial] resources for the Tumen Program have been relatively modest. UNDP has provided approximately US$7 million, and another US$7 million has been contributed by other UN agencies, the Global Environment Facility, and individual countries including South Korea and the Nordic countries' (Winder, 2000, p. 8).
13. 'In relation to this, since late 1993, North Korea built a fence of 3.6 m in height and 80 km in length around the [750 km^2 Najin–Sonbong FEZ] to cut it off from the domestic economy' (Namkoong, 1999, p. 6).

References

Akamatsu, K. (1961), 'A theory of unbalanced growth in the world economy', *Weltwirtschaftliches Archiv*, **86**, 196–217.

Andersson, A. (1981), 'Growth and stagnation of economies with public goods: a neoclassical analysis', in W. Buhr and P. Friedrich (eds), *Lectures on Regional Stagnation*, Baden-Baden: Nomos, pp. 31–49.

Bartik, T. (1991), *Who Benefits from State and Local Economic Development Policies*, Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Batey, P. and P. Friedrich (eds) (2000), *Regional Competition*, Heidelberg: Springer.

Beckmann, M. (1981), 'Are the big cities getting bigger and the small getting smaller? – or how stable is a central place system?', in W. Buhr and P. Friedrich (eds), *Regional Development under Stagnation*, Baden-Baden: Nomos, pp. 219–26.

Beneton, A. (2002), 'Public and welfare economics under monopolistic and competitive governments', in St. L. Winer and H. Shibata (eds), *Political Economy and Public Finance*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, pp. 31–44.

Biermann, H. and P. Friedrich (1972), 'Entwicklung einer gemeindespezifischen Produktionsfunktion', *Environment and Planning*, **4**, 273–321.

Chaudhuri, T.D. and S. Adhikari (1993), 'The locational choice for free-trade zones: rural versus urban options', *Journal of Development Studies*, **41**, 157–62.

Cotton, J. (1994), 'Signs of change in North Korea', *Pacific Review*, **7**, 223–7.

Devereux, J. and L.I. Chen (1995), 'Export zones and welfare: another looks', *Oxford Economic Papers*, **47**, 705–13.

Felderer, B. (1981), 'A theory of stagnation', in W. Buhr and P. Friedrich (eds), *Regional Development under Stagnation*, Baden-Baden: Nomos, pp. 13–30.

Feng, X. and P. Friedrich (1995), 'Ansätze zu einer Theorie des Vertragsmanagements', *Zeitschrift für öffentliche und gemeinnützige Unternehmen*, **18**, 277–97.

Feng, X. and Q. Yang (2007), 'Raumstrukturelle Effekte des Bevölkerungsrückgangs', in X. Feng and A. Popescu (eds), *Infrastruktur und Bevölkerungsrückgang*, Berlin: Berliner Wissenschaftsverlag, pp. 47–62.

Fisher, P.S. and A.H. Peters (1997), 'Tax and spending incentives and enterprise zones', *New England Economic Review*, March/April, 109–30.

Frey, B. (1968), 'Eine politische Theorie des wirtschaftlichen Wachstums', *Kyklos*, **21**, 70–99.

Friedrich, P. (1983), 'Regional aspects of X-inefficiency in the public sector', in E. Wille (ed.), *Konzeptionelle Probleme öffentlicher Planung*, Bern: Lang, pp. 189–244.

Friedrich, P. (1987), 'Regional competition under stagnation', in P. Friedrich and P. van Rompuy (eds), *Fiscal Decentralization*, Baden-Baden: Nomos, pp. 85–133.

Friedrich, P. (1991), 'Prinzipien der Besteuerung für duale sozialistische Wirtschaftsordnungen unter Berücksichtigung von Joint-Ventures', *Jahrbuch für Sozialwissenschaft*, **41**, 312–42.

Friedrich, P. and X. Feng (2002), 'The role of public institutions in regional competition', in G. Atalik and M. Fisher (eds), *Regional Development Reconsidered*, Berlin, Heidelberg: Springer, pp. 79–113.

Ge, W. (1995), 'The urban enterprise zone', *Journal of Regional Science*, **35**, 217–31.

Glaeser, E.L., H.D. Kallal, A. Scheinkman and A. Schleifer (1992), 'Growth in cities', *Journal of Political Economy*, **100**, 1226–52.

Goulet, D. (1971), *The Cruel Choice: A New Concept on the Theory of Development*, New York: Atheneum.

Hamada, K. (1974), 'An economic analysis of the duty-free zone', *Journal of International Economics*, **4**, 225–41.
Hamilton, C. and L.E.O. Svensson (1982), 'On the welfare effects of a duty-free zone', *Journal of International Economics*, **13**, 45–64.
Hansen, H. (1939), 'Economic progress and declining population growth', *American Economic Review*, **29**, 1–15.
Hirschman, A. (1958), *Strategy of Economic Development*, New Haven, CT: Yale University Press.
Hoffmann, W.G. (1955), *The British Industry 1700–1950*, Oxford: Basil Blackwell.
Hoffmann, W.G. (1965), *Das Wachstum der deutschen Wirtschaft seit der Mitte des 19. Jahrhunderts*, Berlin, Heidelberg: Springer.
Hughes, C.W. (2000), '*Tumen River Area Development Programme: frustrated micro-regionalism as a microcosm of political rivalries*', CSGR Working Paper 57/00, University of Warwick.
Ihlanfeldt, K.R. (1995), 'Ten principles for state tax incentives', *Economic Development Quarterly*, **9**, 339–55.
Jung, E.S., Y. Kim and T. Kobayashi (2004), 'North Korea's special economic zones: obstacles and opportunities', in The Korea Economic Institute (ed.), *Beyond the North Korean Nuclear Crisis*, Seoul: The Korea Economic Institute, pp. 43–59.
Kaniss, Ph. (1981), 'The role of regional decline in adaptive transformation', in W. Buhr and P. Friedrich (eds), *Regional Development under Stagnation*, Baden-Baden: Nomos, pp. 77–89.
Kim, S.J. (1996), 'The Tumen River Area Development Program: the present status and future prospects', *The Economics of Korean Unification*, **1**, 105–19.
Kojima, K. (2000), 'The flying-geese model of Asian economic development: origin, theoretical extensions and regional policy implications', *Journal of Asian Economics*, **11**, 375–401.
Kuznets, S. (1966), *Modern Economic Growth, Rate, Structure, and Spread*, New Haven, CT: Yale University Press.
Lim, H.C. and Y.C. Chung (2004), 'Is North Korea moving toward a market economy?', *Korea Focus*, **12**(4), 49–79.
Lindemann, St. (1999), *Theorie und Empirie kommunalen Wirtschaftsförderungswettbewerbs*, Baden-Baden: Nomos.
Meier, G. and R.E. Baldwin (1957), *Economic Development, Theory, History, Policy*, New York: John Wiley & Sons.
Mills, E.S. and F. McDonald (1992), *Sources of Metropolitan Growth*, New Brunswick, NJ: Center for Urban Policy Research of the Rutgers University.
Musleh-ud, D. (1994), 'Export processing zones and backward linkages', *Journal of Development Economics*, **43**, 369–85.
Myrdal, G. (1957), *Economic Theory and Underdeveloped Regions*, London: Duckworth.
Nam, C.W. (2006), 'Development stage theory and industrial growth patterns, Asian NIEs and selected advanced economies compared (1980–1995)', *International Quarterly for Asian Studies*, **37**, 357–94.
Nam, C.W. and D.M. Radulescu (2004), 'Do corporate tax concessions really matter for the success of free economic zones?', *Economics of Planning*, **37**, 99–123.
Namkoong, Y. (1999), 'North Korean external economic policies and inter-Korean economic cooperation', paper presented at the Annual Conference organised by the Political Science Association, University of Nottingham, 23–25 March.
Neubauer, G. (2007), 'Auswirkungen der demographischen Veränderungen auf die Gesundheitsversorgung in Deutschland', in X. Feng and A. Popescu (eds), *Infrastruktur und Bevölkerungsrückgang*, Berlin: Berliner Wissenschaftsverlag, pp. 230–55.
Niskanen W. (1971), *Bureaucracy and Representative Government*, Chicago, IL: Aldine-Atherton.
North, D.C. (1993), 'The ultimate sources of economic growth', in A. Szirmai, B. van Ark and D. Pilat (eds), *Explaining Economic Growth*, Amsterdam: North-Holland, pp. 65–77.
Papke, L.E. (1992), 'What do we know about enterprise zones?', NBER Working Paper, 4251.
Park, J.D. (1996), 'The state and prospects of the Rajin-Sonbong free economic and trade zone', *Economics of Korean Unification*, **1**, 12–17.
Pike, A., A. Rodríguez-Pose and J. Tomaney (2006), *Local and Regional Development*, London: Routledge.
Ranis, G. (2004), 'The evolution of development thinking: theory and policy', Economic Growth Center Discussion Paper 886, Yale University.
Ritschl, H. (1927), 'Reine und historische Dynamik des Standortes der Erzeugungszweige', *Schmollers Jahrbuch*, **53**, 813–70.
Rodríguez, C.A. (1976), 'A note on the economics of the duty free zone', *Journal of International Economics*, **6**, 385–8.
Roscher, W. (1861), *Ansichten der Volkswirtschaft aus dem geschichtlichen Standpunkte*, Leipzig: Winter.
Roscher, W. (1882), *Principles of Political Economy*, translated by J.T. Lalor, Chicago, IL: Callaghan & Company.
Rostow, W. (1960), *The Stage of Economic Growth: A Non-Communist Manifesto*, Cambridge: Cambridge University Press.
Schäffle, A.E.F. (1873), *Das gesellschaftliche System der menschlichen Wirtschaft*, Vol. 2, Tübingen: Mohr.

Schweinberger, A.G. (2003), 'Special economic zones in developing and/or transition economies: a policy proposal', *Review of International Economics*, **11**, 619–29.
Sen, A.K. (1999), *Development as Freedom*, London: Macmillan.
Seston, W. (1963), 'Verfall des römischen Reiches im Westen. Die Völkerwanderung', in G. Mann, A. Heuss and A. Nitschke (eds), *Propyläen Weltgeschichte*, Vol. 4, Frankfurt: Ullstein, pp. 487–604.
Spencer, H. (1882), *Principles of Sociology*, London: William & Norgate.
Spengler, O. (1947), *The Decline of the West*, New York: Alfred A. Knopf.
Szirmai, A. (2005), *Socio-Economic Development*, Cambridge: Cambridge University Press.
Tahir, J. (1999), 'An assessment of free economic zones in Arab countries: performance and main features', Economic Research Forum (for the Arab Countries, Iran and Turkey) Working Paper 9926.
Timm, H. (1963), 'Staat, Wachstum, Preisniveau', *Zeitschrift für die gesamte Staatswissenschaft*, **119**, 253–81.
Tönnies, F. (1887), *Gemeinschaft und Gesellschaft*, Leipzig: Fues's Verlag.
Tuppen, J. (1993), 'Enterprise zones in France: developments and impacts', *Regional Studies*, **27**, 260–64.
United Nations Centre on Transnational Corporations (UNCTC) (1991), *The Role of Free Economic Zones in the USSR and Eastern Europe*, UNCTC Current Studies, Series A, 14, New York: United Nations Publications.
United Nations Conference on Trade and Development (UNCTAD) (1995), *Trade and Development Report 1995*, New York and Geneva: United Nations Publications.
United Nations Conference on Trade and Development (UNCTAD) (1996), *Trade and Development Report 1996*, New York and Geneva: United Nations Publications.
United Nations Conference on Trade and Development (UNCTAD) (2000), *Tax Incentives and Foreign Investment: A Global Survey*, Geneva: United Nations Publications.
United Nations Development Programme (UNDP) (2001), *Human Development Report 2000*, New York: United Nations Publications.
Van Rompuy, P. (1981), 'Allocation and redistribution under reduced growth', in W. Buhr and P. Friedrich (eds), *Lectures on Regional Stagnation*, Baden-Baden: Nomos, pp. 50–69.
Wall, D. (1993), 'China's economic reform and opening-up process: the role of the special economic zones', *Development Policy Review*, **11**, 243–60.
Wallerstein, E. (1974), *The Modern World System*, New York and Sydney: Academic Press.
Westlund, H. (2006), *Social Capital in the Knowledge Economy, Theory and Empirics*, Heidelberg: Springer.
Winder, J.A.B. (2000), 'The economic dynamics of the Korean peninsula peace process', revised paper presented at the Conference on the Korean Peninsula, Northwestern University, 26 May.
World Bank (1997), *World Development Report 1997: The State in a Changing World*, New York: Oxford University Press.
World Bank (2000), *World Development Report 2000/2001: Attacking Poverty*, New York: Oxford University Press.
Young, L. (1987), 'Intermediate goods and the formation of duty-free zones', *Journal of Development Economics*, **25**, 369–84.

Index