Dynamic Modeling and Econometrics in Economics and Finance 19

Pasquale Commendatore
Saime Kayam
Ingrid Kubin *Editors*

Complexity and Geographical Economics

Topics and Tools

Springer

Dynamic Modeling and Econometrics in Economics and Finance

Volume 19

Editors
Stefan Mittnik
Ludwig Maximillian University Munich
Munich, Germany

Willi Semmler
Bielefeld University
Bielefeld, Germany
and
New School for Social Research
New York, USA

More information about this series at
http://www.springer.com/series/5859

Pasquale Commendatore • Saime Kayam •
Ingrid Kubin
Editors

Complexity and Geographical Economics

Topics and Tools

Springer

Editors
Pasquale Commendatore
Department of Law
University of Naples "Federico II"
Naples, Italy

Saime Kayam
Department of Management Engineering
Istanbul Technical University
Macka, Istanbul
Turkey

Ingrid Kubin
Department of Economics
WU Vienna University of Economics and Business
Vienna, Austria

ISSN 1566-0419 ISSN 2363-8370 (electronic)
Dynamic Modeling and Econometrics in Economics and Finance
ISBN 978-3-319-12804-7 ISBN 978-3-319-12805-4 (eBook)
DOI 10.1007/978-3-319-12805-4

Library of Congress Control Number: 2015932772

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

The uneven geographical distribution of economic activities is a huge worldwide challenge. For the European Union (EU) regions, this is shown by the deep differences within and across nations. Spatial inequalities are evolving through time following complex patterns determined by economic, geographical, institutional and social factors. The New Economic Geography approach, which was initiated by P. Krugman in the early 1990s, describes economic systems as very simplified spatial structures. Developing a more sophisticated modelling of the EU—visualized as an evolving trade network with a specific topology determined by the number and strength of national, regional and local links—can provide economic policies specifically designed to take into account this pervasive network structure assessing the position of backward locations within the network and focussing on instruments that favour interconnections. The ISCH COST Action IS1104 "The EU in the new complex geography of economic systems: models, tools and policy evaluation" approved by the European Union in 2011 and funded for the 4-year period 2012–2016 is a network connecting more than 80 researchers in 25 European countries devoted to this task. The Action results are expected to provide a basis for an improved evaluation of such policies, in particular for the European Cohesion Policy, considering their impact on the welfare level of EU citizens and its geographical distribution. To achieve this objective, the Action enhances interdisciplinary networking combining recent approaches in economics with the most advanced mathematical, empirical and computational methods for analysing complex and non-linear systems.

This book, which is mainly a collection of literature reviews on relevant aspects of the Action, is put together to provide a basis for further analysis and as an introduction into the complex network approach to the European Regional Policy. The book consists of three main parts, i.e. economic geography modelling, institutions and markets, and social and industrial interactions. Each of these parts actually overlaps with separated but complementary lines of research followed in understanding the interlinkages and interactions that emerge at different levels involving different territorial units, institutions and individual agents.

The prospective readers are expected to be academicians and researchers who demand new tools to pursue their investigations in the field. The contributors are all experts on either topics or tools proposed to examine economic geography and networks and are members of the Action. The contributions in the present volume were presented and discussed in a number of workshops and conferences organized within the Action in 2012 and 2013.

Naples, Italy — P. Commendatore
Istanbul, Turkey — S. Kayam
Vienna, Austria — I. Kubin

Acknowledgements

We should first acknowledge the help of the organizers of Action workshops and conferences in Urbino (18–19 September 2012), Lisbon (25–28 April 2013), Naples (3 May 2013), Madrid (20 May 2013) and Siena (4–6 July 2013). They have provided the necessary environment for fruitful discussions for the experts and members of the Action without whom neither the conception nor the contents of this book could have been achieved. Likewise, we are also grateful to the contributors for their timely and patient responses to our comments in the editorial and production processes.

Our special thanks go to the external referees, who not only reviewed the chapters allocated to them but also made significant contributions to improve the contents. The editors are grateful to Raffaello Bronzini (Bank of Italy), Julie Le Gallo (CRESE - Université de Franche-Comté) and the Action member Jose S. Cánovas (Politécnica de Cartagena) for doing an excellent and timely work as referees.

We are grateful to the series editors Stefan Mittnik and Willi Semmler, and we would like to give our appreciation to Martina Bihn, Ruth Milewski, Yuliya Zeh and Venkatachalam Anand from Springer for their patient and constructive attendance during the editorial process.

Contents

Contributors

Marcel Ausloos Royal Netherlands Academy of Arts and Sciences, Amsterdam, The Netherlands (previously at GRAPES, ULG, Liege, Belgium)

Roberto Basile Department of Economics, Second University of Naples, Capua (CE), Italy

Lutz Bellmann Friedrich-Alexander-Universität Erlangen-Nürnberg and Institut für Arbeitsmarkt-und Berufsforschung, Nürnberg, Germany

Gian Italo Bischi Department of Economics, Society, Politics, University of Urbino "Carlo Bo", Urbino, Italy

Björn Brand Department of Organization and Economics of Institutions, University of Graz, Graz, Austria

Spiros Bougheas School of Economics, University of Nottingham, Nottingham, UK

Jesús Carrete Departamento de Física da Materia Condensada, Univ. de Santiago de Compostela, Santiago de Compostela, Spain
LITEN, CEA-Grenoble, Grenoble, France

Pasquale Commendatore Department of Law, University of Naples "Federico II", Naples, Italy

Herbert Dawid Department of Business Administration and Economics and Center for Mathematical Economics, Bielefeld University, Bielefeld, Germany

Valerio Filoso Department of Law, University of Naples "Federico II", Naples, Italy

Hans-Dieter Gerner Institut für Arbeitsmarkt-und Berufsforschung, Nürnberg, Germany

Theresa Grafeneder-Weissteiner Department of Economics, University of Vienna, Vienna, Austria

Saime Kayam Department of Management Engineering, Istanbul Technical University, Istanbul, Turkey

Alan Kirman Faculte de Droit et de Sciences Politiques, Directeur d'études à l'EHESS, Membre de l'IUF, GREQAM, Aix-Marseille Université, Aix-en-Provence, France

Michael Kopel Department of Organization and Economics of Institutions, University of Graz, Graz, Austria

Ingrid Kubin Department of Economics, Vienna University of Economics and Business, Vienna, Austria

Fabio Lamantia Department of Economics, Statistics and Finance, University of Calabria, Rende (CS), Italy

Diana A. Mendes Department of Quantitative Methods for Business and Economics, ISCTE-IUL and BRU-IUL, Lisbon, Portugal

Vivaldo M. Mendes Department of Economics, ISCTE-IUL and BRU-IUL, Lisbon, Portugal

Ugo Merlone Department of Psychology, Università di Torino, Torino, Italy

Román Mínguez Statistics Department, University of Castilla-La Mancha, Cuenca, Spain

Jose María Montero Statistics Department, University of Castilla-La Mancha, Cuenca, Spain

Jesús Mur Department of Economic Analysis, University of Zaragoza, Zaragoza, Spain

Douglas R. Nelson Tulane University, Murphy Institute, New Orleans, LA, USA

Anastasiia Panchuk Institute of Mathematics NAS of Ukraine, Kiev, Ukraine

Mario Pezzino School of Social Sciences, University of Manchester, Manchester, UK

Tönu Puu CERUM, Umeå University, Umeå, Sweden

Giulia Rotundo Department of Methods and Models for Economics, Territory and Finance, Sapienza University of Rome, Rome, Italy

Richard Upward University of Nottingham, School of Economics, Nottingham, UK

Luis M. Varela Grupo de Nanomateriais e Materia Branda, Departamento de Física da Materia Condensada, Univ. de Santiago de Compostela, Santiago de Compostela, Spain

The Editors

Pasquale Commendatore is Full Professor of Economics at the University of Naples 'Federico II', Italy. His fields of interest are: Geographical Economics, Economic Dynamics, Development Economics and Economic Growth. His works are published in international journals such as *Oxford Economic Papers; Cambridge Journal of Economics; Nonlinear Dynamics, Psychology, and Life Sciences; Journal of Economic Behavior and Organization; Spatial Economic Analysis; Computational Economics and Structural Change and Economic Dynamics*. From 2007 he is Member of the Scientific Board of CICSE (Centro Interuniversitario Crescita e Sviluppo Economico—Interuniversity Centre on Growth and Economic Development). He is the proposer and elected Chair of the COST (European Cooperation in Science and Technology) Action IS1104 "The EU in the new complex geography of economic systems: models, tools and policy evaluation", a research project supported by the EU RTD Framework Programme (2012–2016).

Saime Kayam is Associate Professor of International Economics at Istanbul Technical University, Turkey. Her fields of interest are: International Trade, Foreign Direct Investments, Regional Economics and Geographical Economics. Her works on multinational corporations and spatial distribution of foreign direct investments have been published as papers in international journals such as *Transition Studies Review of Economics, Business and Finance* and as chapters in contributed volumes. She mainly focuses on dispersion of economic activity in emerging and developing countries and has several studies on outward investments from these group of countries. Dr. Kayam is the Vice-Director of ESRC-ITU (the Economic and Social Research Centre) at Istanbul Technical University. She is a member of the Scientific Committee of the Action.

Ingrid Kubin is Full Professor of International Economics at WU, Vienna University of Economic and Business, Austria. Her fields of interest are: Geographical Economics, Economic Dynamics and International Economics. Her works are published in international journals such as *Mathematics and Computers in Simulation; Journal of Evolutionary Economics; Economia Politica; Computational Economics; Oxford Economic Papers; Metroeconomica; Spatial Economic*

Analysis; Economie Internationale; Nonlinear Dynamics, Psychology, and Life Sciences; Journal of Economic Behavior and Organization; Economics Letters and International Regional Science Review. She is on the editorial board of *Empirica—Journal of European Economics* and is elected Vice-Chair of the COST Action IS1104 "The EU in the new complex geography of economic systems: models, tools and policy evaluation".

Introduction

Pasquale Commendatore, Saime Kayam, and Ingrid Kubin

Abstract The uneven geographical distribution of economic activities is a huge worldwide challenge. Spatial inequalities are evolving through time following complex patterns determined by economic, geographical, institutional and social factors. The New Economic Geography approach, which was initiated by P. Krugman in the early 1990s, describes economic systems as very simplified spatial structures. This book aims at providing an overview of the existing state of knowledge and new perspectives of research to set the basis for developing a more sophisticated modelling of the economic activities visualised as being influenced by evolving trade networks with a specific topology that is determined by the number and strength of national, regional and local links. To achieve this objective the chapters combine recent approaches in economics with the most advanced mathematical and computational methods for analysing complex and non-linear systems to build an interdisciplinary understanding of the issue.

The problem of uneven geographical distribution of economic activities is a huge worldwide challenge. For the European Union (EU) regions this is shown by the deep differences within and across nations. According to the Eurostat regional yearbook 2013, the GDP per inhabitant of 41 EU-27 NUTS 2 regions, out of 270, is above 125 % of the average, whereas that of 68 regions falls below 75 % of the average; 25 of those "below average" regions are found in six of the EU-15 member

P. Commendatore (✉)
Department of Law, University of Naples "Federico II", Via Mezzocannone 16, 80138 Naples, Italy
e-mail: pasquale.commendatore@unina.it

S. Kayam
Department of Management Engineering, Istanbul Technical University, Suleyman Seba C. No. 90, Macka, Istanbul, Turkey
e-mail: kayams@itu.edu.tr

I. Kubin
Department of Economics, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria
e-mail: Ingrid.kubin@wu.ac.at

P. Commendatore et al. (eds.), *Complexity and Geographical Economics*,
Dynamic Modeling and Econometrics in Economics and Finance 19,
DOI 10.1007/978-3-319-12805-4_1

States (Greece, Italy, France, Portugal, United Kingdom and Spain). The remaining 43 regions are in Member States that joined the EU between 2003 and 2007; 22 of which have an average GDP even below 50 % of the EU-27 average (these are found in Bulgaria, Hungary, Poland, Romania and Slovakia).

Regional disparities are significant not only across but also within countries (especially Turkey, Slovakia, but also Italy, Germany and the UK) due to historical, economic, financial, political, geographical, institutional and social factors. Spatial inequalities evolve through time. For the period 2000–2007 growth trends in the EU show some overall catching-up at the country level but with remarkable differences at the regional level. The economic crisis in 2008–2009 has slowed down the convergence process hitting some regions and nations more than others (IMF 2011 Regional Economic Outlook: Europe).

The complex nature of the distribution of economic activity across space and of its evolution through time requires necessarily different levels and tools of analysis. For the current volume, we differentiate between a macro—bird's eye—perspective looking at regions and countries; a meso perspective looking at markets and institutions; and a micro perspective looking at single agents, in particular at producers. For each perspective, we provide reviews of the economic theory and of the analytic and empirical methods necessary for the respective levels of disaggregation.

1 Economic Geography Modelling

Krugman (1993)—with reference to Cronon's famous book "Nature's Metropolis: Chicago and the Great West" (Cronon 2009)—popularized the distinction between "first nature" and "second nature" in explaining why economic activity is not evenly distributed over space. First nature factors are exogenous to economic activity, such as endowment with natural resources (freely available as sunshine or other climatic features; or costly available such as minerals or coal), and geographic and geopolitical factors (location by the sea or river, geographic centrality, institutional or political differences). Traditional models of international commodity trade (such as the Heckscher–Ohlin model) and of international factor migration would predict for open economies an uneven distribution of economic activity that reflects the distribution of first nature factors.

More recently, regional growth models are also widely used in studying spatial processes of convergence or divergence. Also in this perspective, first nature differences—reinforced by local knowledge spillovers—mitigate regional convergence processes thus leading to an uneven distribution of economic activity.

However, the main focus of the present volume is on models of the New Economic Geography (NEG); its main achievement is to show that even with first nature symmetric regions the formation of Core-Periphery structures, in which economic activity is concentrated in one of the regions, is possible via endogenous (second nature) agglomeration processes that involve human activity and economic

incentives. In typical NEG models, the role of first nature is minimal. Economic activities tend to cluster together taking advantage of proximity to larger markets, scale economies, knowledge spillovers, lower transport costs etc. Even starting from undifferentiated regions and large dispersion once the process is set in motion all activities tend to agglomerate. A fortiori, the concentration process is favoured by first nature causes determined by territorial topography, endowments of natural resources and geopolitical factors. Indeed EU core regions are typically more urbanised (capital regions are often the richest) and well-connected to the transport network hubs whereas peripheral regions are often coastal, on borders or in rural areas, with a lower number of connections. The increase in the number and strength of territorial connections (transport networks and so on) is at the centre of many EU policies aiming to intensify economic, social and territorial cohesion and to enhance competitiveness (Fifth European Commission Report on economic, social and territorial cohesion).

From a modelling perspective, geographical space can be represented in discrete form as a set or matrix of locations connected by arcs or as a continuous plane in two dimensions. Economic activities taking place in one location, such as production and consumption, and flows connecting two locations, such as commodity trade and labour migration, can be represented in both forms leading to two alternative modelling strategies in geographical economics. The continuous specification has been the tradition both in economics and geography. However, it does not as easily lend itself to computations and fitting to actual data that are a discrete representation. More recently, the New Economic Geography uses primarily a discrete representation of geographical space, with the danger that space, in the sense of geometric shape and size, slips out.

The chapter by **Commendatore, Filoso, Grafeneder-Weissteiner and Kubin** reviews the NEG models that are formulated in discrete space and that explain the uneven distribution of economic activity across space employing endogenous agglomeration processes, and the interplay of agglomeration and dispersion forces. Unlike the two-region models that have a restricted applicability, this chapter focuses on the multiregional NEG models recently developed and reviews their contributions systematically. As unifying framework, the authors use the footloose entrepreneur model considering the role and nature of transport costs, and the welfare properties of the spatial equilibria and the design of optimal economic policies.

In his chapter **Puu** presents a detailed study of diffusion processes for some classical macroeconomic dynamic models in continuous geographical space; in doing so, the contribution also reviews the involved mathematical tools. More specifically, the author applies the so-called Laplacian operator to a Harrodian growth model of an open economy and to a non-linear business cycle model based on the classical accelerator mechanism. With this approach it is possible to show the spatial pattern of growth and decline—considering the projection on a horizontal planar space of the corresponding linear business cycle model—for the growth model; and the cyclical motion of an economic system represented as overlapping

shrinking flat surfaces in a three dimensional space for the nonlinear business cycle model.

The next two chapters that are by Mendes and Mendes, and by Basile, Kayam, Mínguez, Montero and Mur review pertinent empirical methods; in particular spatial econometric methods. In their contribution, **Mendes and Mendes** present *parametric* spatial econometric models that can be applied to regional economics. They introduce and discuss the basic terminology, the spatial data dependence, the specification of spatial effects, and some basic spatial regression models, i.e. the spatial autoregressive model (SAR), the spatial error model (SEM), the spatial Durbin model and two general spatial models. The maximum likelihood estimation for SAR and SEM models is also presented in detail. Finally, the contribution presents several empirical works in the context of the European Union focusing on particular areas of urban economics, economic growth and productivity, and studies dealing with agglomeration and externalities spillovers).

The other chapter on spatial econometrics by **Basile, Kayam, Mínguez, Montero and Mur** argues that modelling regional economic dynamics requires the adoption of complex econometric tools, which allow to deal with some important methodological issues arising in a spatial context, such as spatial dependence, spatial heterogeneity and nonlinearities; the authors argue that *semiparametric* approaches in the spatial econometrics literature have recently provided some instruments, which address these issues simultaneously. In particular, the authors present a spatial autoregressive semiparametric geoadditive (SAR-Geo-AM) model and discuss technical issues concerning its estimation (including the use of restricted maximum likelihood methods that allow estimating the parameters of SAR-Geo-AM model in a single step). Finally, the authors review the empirical literature on regional economic dynamics and economic geography and present, in particular, an application of the SAR-Geo-AM to regional growth data.

2 Institutions and Markets

A closer look at different territorial levels (EU, national, regional and local) reveals that countries and regions are interconnected by various networks and that these networks play a crucial role in fostering growth and reducing regional disparities. Their functioning is shaped by market (but also nonmarket) institutional setups; more specifically, we focus on networks of financial markets and of labour markets.

Public and private finance are deeply interconnected across space, as the recent waves of financial turmoil have revealed, and can be described as a global network (with also a European scale) within which a few large centres assume a central position. At the same time these networks have also a local dimension, which stresses informational advantages enjoyed by financial intermediaries located in the proximity of firms. Over time, the networks evolve, some centres assuming higher importance whereas others fall back. The different aggregation levels are also interconnected and institutional reforms change their respective weights. For public

finance, we observe at the EU scale a growing burden of sovereign debts and large national budget deficits that may endanger economic and financial integration and that call for better policy coordination among national governments and monetary authorities. At the national level, as a consequence of the economic crisis, the share in national GDPs of public expenditure is rising. However, a marked process of decentralisation in public expenditure has taken place in the last few years (for example, two thirds of public investment is carried out, on average in the EU, by sub-national governments, regional or local) and there are categories of public expenditure that have a strong local impact (such as transport infrastructure and environmental policies). Institutional capacities however are unevenly distributed across space. Improving the quality of governance (at the various levels) and developing better linkages and coordination between central and local governments and among local administrations become strategic issues.

In their chapter, **Bougheas and Kirman** take a closer look at the network of financial institutions interconnected by financial exposure or by financial transactions. They review papers that describe the network structure for banking systems in various countries; and elaborate the relationship between the structure of the network, its topological properties and the propagation of an external shock (and thus the fragility of the entire system). The main focus of the review is on the use of complex network analysis techniques and their applications for analysing and pricing the systemic risk of a financial network.

Also labour markets have a network dimension and a spatial dimension. They are connected via a migration network and institutions heavily shape the functioning of labour markets and of the migration network. They are regionally fragmented because labour mobility is low as a consequence of institutional factors, such as language, territorial, cultural, gender, ethnic, age and other barriers across European communities.

The chapter by **Nelson** provides an overview of current research on networks in international migration. It begins with a short discussion of the relationship between networks and social capital. While controversial, this concept potentially provides a unifying thread linking various aspects of economic research and, potentially more importantly, providing a bridge linking economic research to parallel research in demography and sociology. The core of the chapter discusses the role of networks in the decision to migrate, the role of networks in assimilation, and the effect of global migrant networks on the pattern of international trade. In all three of these areas, recent years have seen substantial new research, both theoretical and empirical, on the ways networks interact with more standard economic variables. In each of these cases, networks are seen to play an essential role in the migration experience.

The chapter by **Bellmann, Gerner and Upward** "zooms in" on one particular labour market and investigates the role of institutions. It starts from the observation that the global economic and financial crisis, which began in 2008, had very different effects on the labour markets of EU economies; in particular, the German labour market might be described as more "resilient" than others in the face of shocks. In this chapter the authors propose a simple descriptive methodology that allows to shed light on many of the proposed explanations for the resilience

of the German labour market to the crisis, in particular on the role of various institutional arrangements intended to promote workplace flexibility, such as short-time-work and working time accounts. The paper focuses on Germany; however, the contribution also describes the used methodology in detail so that it can be consistently applied across countries (given that detailed linked employer-employee data are increasingly available)

Finally, **Varela, Rotundo, Ausloos and Carrete** provide a brief introduction to complex network analysis including computational issues. After an introduction to the foundations of the field, the authors add insights on the statistical mechanical approach, and on the most relevant computational aspects for the treatment of these systems. As it is the most frequently used model for interacting agent-based systems (that are often used in economics), a brief description of the statistical mechanics of the classical Ising model on regular lattices, together with recent extensions of the same model on small-world Watts-Strogatz and scale-free Albert-Barabási complex networks, is included. The authors provide many references for further studying these methods and review applications in the broader field of social sciences—with a special focus on applications in economics (such as business cycle coordination; financial markets, tax evasion, business and innovation networks, international trade and migration networks).

3 Industrial Interactions

Decisions that impact on population's well-being (labour migration movements, households residential choices as well as firms location decisions) may also be affected by lower scale interrelations: the number and strength of social ties contributes to explaining differences in occupational opportunities and wage outcomes; residential choices have a clear spatial dimension (characteristics of local housing markets, accessibility, neighbourhood quality) involving individual preferences but also linked to households interactions; firms compete in local as well as in international markets, where larger firms are involved with more scope for strategic interactions. Our aim in this part is to provide a disaggregated analyses of multi-level spatial economic systems focusing on the interrelationship between individual location choices and the economic, social and institutional environment; to analyze the social and economic networks that may emerge at various scales; and, finally, to identify possible interconnections among networks, to highlight how the properties of a network at some level (for example, the degree of interconnectedness) can be affected by processes taking place within networks at a different level. A secondary, but not unimportant objective is to provide more sophisticated descriptions of agents' behaviour that could suitably replace the oversimplified monopolistically competitive behaviour within NEG models.

Concentrating on the industrial interactions between agents, i.e. firms, we take a micro perspective in examining individual agent's behaviour and small scale interactions and networks. Firms may interact on a cooperative basis creating bilateral or multilateral links (i.e. building the so-called innovation networks, or other types of inter-firm networks). They may decide either to engage in competitive relationships, through various types of strategic behaviour; or create cooperative links such as innovation networks or other types of inter-firm networks. The existence of local links increases substantially the importance of local administration quality and interventions. Firms' location decisions may be affected by industrial interactions.

Starting from the first principles of industrial organization **Panchuk** develops the dynamics of industrial interaction where the competition between firms determines how intensely capital is utilized and the level of capital stock chosen by the firms at the end of each investment period. This chapter investigates how a market structure is developed when several firms are involved. Any new industry, not depending on how large it may expand in the course of time, is originally established through a few pioneering firms, and eventually starts growing in terms of the number of competitors, thus developing competition. It is supposed that the firms act under constant eventually decaying returns, and that they cause in competitors the need to renew their capital equipment from time to time, choosing its optimal amount according to the current market situation. Meanwhile, in the intervening periods the firms are subjected to capacity limits due to fixed capital stocks. As a result, the evolution of the system depends essentially on the number of competitors and the capital lifetime, and is also sensitive to the initial choice of individual inactivity times. In particular, the firms may merge into different groups renewing their capitals simultaneously, which leads to distinct dynamical patterns.

Firms not only interact through competition but also their spatial presence generates spillovers. R&D investments and spatial spillovers are considered in the chapter by **Bischi and Lamantia** that overviews the literature concerning oligopoly models where firms produce homogeneous goods and share R&D cost-reducing results through bilateral agreements and/or involuntary spillovers of knowledge. In these models the industrial interactions are expressed by the formalism of networks (i.e. theory of graphs) where firms represent nodes and agreements to share R&D represent arcs (or links). The authors describe several models and corresponding theoretical results, as well as some of their practical implications in industrial organization. The second part of the chapter describes a recent dynamic two-stage model of oligopolies with both R&D collaboration ties and spillover effects.

The following chapter by **Kopel, Pezzino and Brand** reviews the theoretical approaches employed to analyze a firm's strategic location choice in an oligopolistic environment by considering its investments and activities regarding R&D. They focus on a firm's sourcing channel choice and examine firms' strategic interaction by means of the analysis of a firm decision's influence on its competitors' strategies. They show the significance of a firm's strategic motives during its decision-making process for both itself and its rivals.

Location choice decision is determined not only through strategic interactions but also through other micro and macro factors. The chapter by **Basile and Kayam** examines the empirical methods employed in analysing the foreign firms' location decisions based on the theoretical literature on macro and micro perspectives of location choice. Starting from the most influential theoretical contributions, which have addressed the motivation of MNEs to be engaged in a horizontal or a vertical FDI, they discuss the various econometric specifications used in the empirical literature to test the hypotheses on the determinants of foreign firms' location. They provide a critical assessment of empirical approaches and their contributions to our understanding of the dispersion of multinational activities across space. Additionally, issues for further development, specifically for modelling multinationals' economic activity in space, are discussed.

Dynamics of geographic distribution of economic activities, including firms' location decision, cause space to become a key element in establishing interactions between individual agents. **Ausloos, Dawid and Merlone** emphasize that understanding of patterns emerging from such spatial interaction between agents is a key problem as much as their description through analytical or simulation means. They employ Agent Based Modelling (ABM) that has become a widespread approach to model complex interactions where agents can interact either indirectly through a shared environment or directly with each other. In such an approach, higher-order variables such as commodity prices, population dynamics or even institutions, are not exogenously specified but instead are seen as the results of interactions. The chapter reviews different approaches for modelling agents' behavior, taking into account either explicit spatial (lattice based) structures or networks. Some emphasis is placed on recent ABM as applied to the description of the dynamics of the geographical distribution of economic activities,—out of equilibrium. The Eurace@ Unibi Model, an agent-based macroeconomic model with spatial structure, is used to illustrate the potential of such an approach for spatial policy analysis.

4 Final Remarks

We believe that the Chapters included in this Volume provide a useful overview of models and tools dealing with multiregional and spatial economies pointing out problems or issues for further development, such as asymmetric regions, multilevel network structures and interactive and strategic behavior leading to location decisions. The book integrates research perspectives on this topic across different disciplines, in particular it integrates specialists in economics and in regional science with specialists in mathematical and computational methods for analyzing complexity and nonlinear dynamics. This interdisciplinary approach allows trespassing the narrow limits set by conventional analytical methods and will allow deriving results where conventional methods have reached their limits.

We expect that this Volume will contribute to theoretical developments visualising the EU as a trade network with a specific topology determined by the number and strength of regional links; to the provision of a more sophisticated modelling of the dynamic processes governing the spatial distribution of industrial activities and financial resources; to the development of specific analytical tools in the field of networks analysis, agent based modelling, evolutionary game theory and nonlinear dynamics. We hope this approach to be more effective for addressing specific cogent issues linked to economic integration such as: easing economic disparities within and across European regions; economic and social cohesion policies; regulation of migration flows; counteracting delocation of production towards low-wage and less regulated emerging economies; upgrading of product quality to enhance the competitive strength of European industries; containing the spread across regions of the consequences of financial markets turbulence.

References

Cronon, W. (2009). *Nature's metropolis: Chicago and the Great West*. WW Norton & Company.

Krugman, P. (1993). First nature, second nature, and metropolitan location. *Journal of Regional Science, 33*(2), 129–144.

Part I
Economic Geography Modelling

Towards a Multiregional NEG Framework: Comparing Alternative Modelling Strategies

Pasquale Commendatore, Valerio Filoso, Theresa Grafeneder-Weissteiner, and Ingrid Kubin

Abstract This chapter reviews the New Economic Geography (NEG) models that explain the uneven distribution of economic activity across space employing endogenous agglomeration processes, and the interplay of agglomeration and dispersion forces. Unlike the two-region models that have a restricted applicability, this chapter focuses on the multiregional NEG models recently developed in the literature, reviews their contributions systematically and compares different modelling strategies. As unifying framework, we use the footloose entrepreneur model considering the role and nature of transport costs, and the welfare properties of the spatial equilibria and the design of optimal economic policies.

1 Introduction

Economic activity in Europe is quite unevenly distributed over the regions: Fig. 1 represents a map of the GDP per head in the European NUTS-3 regions and reveals substantial differences, not only between countries but also between regions of a single country. Why is this the case? This is a core question in geographical economics.

One main approach is the New Economic Geography (NEG), a new paradigm initiated by Krugman (1991). In this perspective, countries are connected via

P. Commendatore (✉) • V. Filoso
Department of Law, University of Naples "Federico II", Via Mezzocannone 16, 80134 Naples, Italy
e-mail: pasquale.commendatore@unina.it; valerio.filoso@unina.it

T. Grafeneder-Weissteiner
Department of Economics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria
e-mail: theresa.grafeneder-weissteiner@univie.ac.at

I. Kubin
Department of Economics, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria
e-mail: Ingrid.kubin@wu.ac.at

P. Commendatore et al. (eds.), *Complexity and Geographical Economics*,
Dynamic Modeling and Econometrics in Economics and Finance 19,
DOI 10.1007/978-3-319-12805-4_2

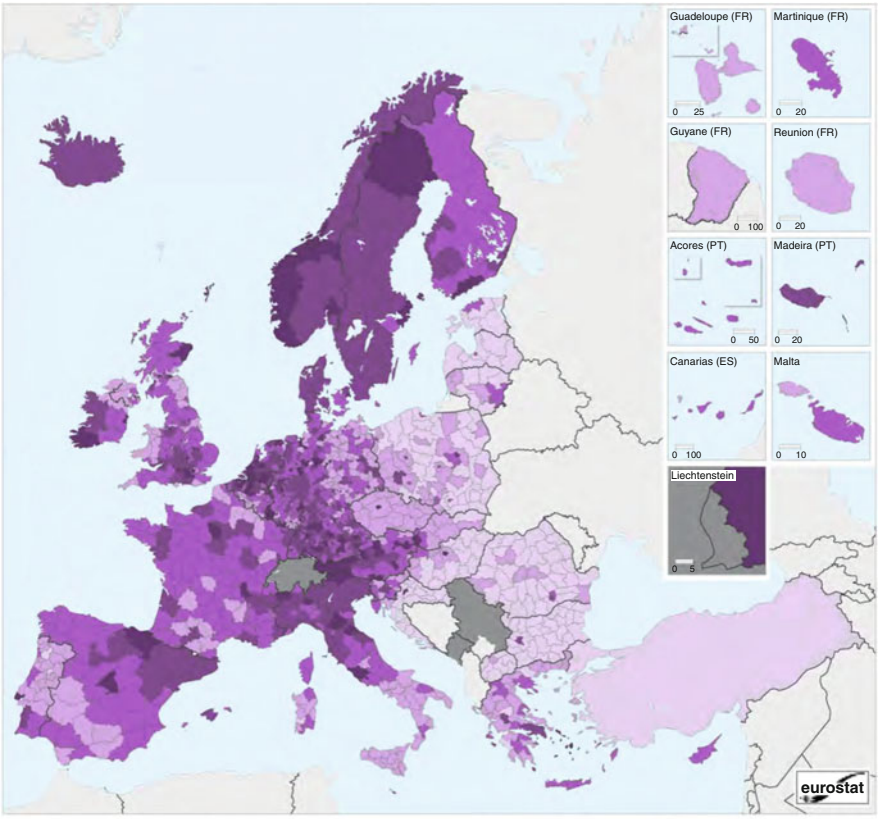


Fig. 1 Gross Domestic Product (GDP) per inhabitant, in purchasing power standard (PPS), by NUTS3 regions, 2010 (% of the EU-27 average, EU-27 = 100). *Source*: Eurostat, Eurostat regional yearbook, 2013

commodity trade and agglomeration is due to a change in the locally available factors of production, in most models—and also in the models we are concentrating on in this review—brought about by factors migrating between the regions

(whereas the total amount of factors is assumed to be constant).[1] A straightforward explanation pattern would recur to an uneven distribution of so-called first nature factors (see also Venables 2006; Roos 2005) that are exogenous to economic activity such as endowment of natural resources (freely available as sunshine or other climatic features; or costly available such as minerals or coal), and geographic and geopolitical factors (location at sea or river, geographic centrality, but also institutional or political differences). These factors determine utility levels either directly or indirectly (via their influence upon production and/or transport costs), thus influencing the location of mobile factors and as a consequence determining geographic patterns of economic activity.

However, the NEG starts with first nature symmetric regions (i.e. with preferences and production technologies that are identical across space) and shows how the formation of Core–Periphery (CP) structures, in which economic activity is concentrated in one of the regions, is still possible via endogenous (second nature) agglomeration processes. In typical NEG models, markets (for manufactured goods) are characterised by Dixit–Stiglitz monopolistic competition (whose main ingredients are preferences that exhibit a taste for variety and decreasing average costs) with isoelastic demand functions. In this framework, factor rewards are higher in the bigger market. In addition, trading of the manufactured goods between the regions involves iceberg transport cost (defining two, or more distinct regions, see Østbye 2010; this is the minimal role of first nature in typical NEG models). Thus, firm location matters. Last, differences in the regional utility levels guide mobility of productive factors between regions. These three elements are found in a wide range of NEG models and are at the heart of Krugman's self-reinforcing agglomeration processes. We exemplify these mechanisms using the footloose entrepreneur model, but a similar interplay of various agglomeration and dispersion forces is found in many NEG models.

In a footloose entrepreneur (FE) model, it is (human/knowledge) capital—embodied in an 'entrepreneur' (or with a slightly different interpretation) in a skilled worker—that is the interregional mobile factor. Thus, firms, (human/knowledge) capital and 'entrepreneurial' (or skilled labour) expenditure shift simultaneously. In a famous thought experiment illustrating the various forces at work, Krugman (1991), considering a two-region economy, asks what happens if one unit of capital is moved from one region to the other—starting from a situation in which capital is equally distributed between the regions. The main agglomerative force (introducing a positive, self reinforcing feedback loop) is the so-called market size effect: with an increase in the local factor amount, the local expenditures and thus the size of the local market increases as well, leading to a higher local factor remuneration which triggers further migration to this region. Since firms are selling to all markets, the size of the local market is less important with lower transport costs and the market size effect is weaker. A price index effect works as additional agglomeration force:

[1] In some other models the same result is obtained via different growth rates of regional factor stocks—as e.g. in the constructed capital model (put forward by Baldwin 1999).

with a higher amount of local productive factors and thus with more local firms more product varieties are locally produced and can be obtained without paying transport costs. Therefore, the price index is lower in the region with a higher amount of productive factors, which leads to further migration. Again, this force is the weaker, the lower the transport costs are. In contrast, the main dispersion force (introducing a negative feedback loop) is the competition effect: with an increase in the local factor amount the number of local firms increases thus leading to more intensive local competition. Local factor remuneration is reduced and migration incentives are weaker. The lower the transport cost, the location of the competitors is less important and the competition effect is weaker. Various NEG models exhibit additional dispersion forces such as immobile demand; congestion effects (e.g. in housing markets), and so on. Typically, it can be shown that for high transport cost dispersion forces prevail leading to an even distribution of economic activity; and that for low transport cost agglomeration forces are stronger resulting in a Core–Periphery pattern of industrial activity.

Notwithstanding the fruitfulness of the NEG approach, the analysis is often limited to the two-region case. As stated recently by Fujita and Thisse (2009), the existence of more than two regions may involve effects that cannot emerge in a two-region context. According to these authors "when there are only two regions, any change in structural parameters necessarily affects directly either one of the two regions, or both. On the contrary, when there are more than two regions, any change in parameters that directly involves only two regions now generates spatial spillover effects that are unlikely to leave the remaining regions unaffected. This in turn further affects the other regions and so on." (Fujita and Thisse 2009, p. 117). When more than two regions are involved, it is much easier to represent and explain spatial configurations of the economic activity that are more similar to those of the real world and that do not emerge in a two-region framework. A typical example is a hub and spoke configuration, in which peripheral regions (the spokes) are reached through a more central location (the hub): in this context, the central position could make that region more attractive or, on the contrary, because of a wider exposure to external competition cause outward firms migration.

In this chapter, we present a general multiregion NEG framework. This framework is a generalization of NEG multiregional models already existing in the literature and it will be used to review such contributions. The review of this literature represents the main objective of the chapter. As any other work of this kind, we need to fix the boundaries of our survey. The distinguishing features of these models other than the common NEG approach (imperfect competition; increasing returns; taste for variety; positive and significant trade costs); are: (a) discontinuous space, that is, we assume a countable and finite number of regions; (b) identical preferences and technologies across space and agents; (c) the size of the overall economy is fixed (no growth); (d) no first nature advantages/disadvantages, like natural resources, better or worse climatic condition, and so on (only two exceptions: the possibility of a more central (peripheral) location; and a larger (or smaller) agricultural sector).

In these models the final outcome of trade liberalization/falling trade costs crucially depends on the modelling strategy. We consider two of them that differ

in the specification of consumer preferences and on trade costs: (1) the standard FE model based on the Dixit–Stiglitz approach and on CES preferences—which is the one adopted by Krugman (1991); (2) the linear FE based on a quadratic utility function and proposed by Ottaviano et al. (2002). We briefly discuss other modelling strategies when presenting our taxonomy.

Given the limited scope of this chapter, we will only sketch the general structure of a multiregional and multi-country (allowing for political boundaries) model and focus mainly on the special case of a 2-country, 3-region model, following the same scheme adopted by Ago et al. (2006) for the case of a 1-country, 3-region model.

Finally, this review will touch two crucial topics: the role and nature of transport costs, which are particularly relevant since they determine the spatial structure of the economy; and the welfare properties of the spatial equilibria and the design of optimal economic policies.

2 Multiregional NEG Models: A Common Framework

In this section we *sketch* a general framework for a multiregional NEG model, i.e. we provide a very preliminary analysis. In particular, we consider a variant of the Footloose Entrepreneur (FE) model—developed by Forslid and Ottaviano (2003). Given the objectives of this chapter, we will not work out its analytical properties in full details (leaving that for future work). However, we discuss specific examples in order to give the reader some hints on how it can be applied and in which directions it could be developed.[2]

2.1 General Framework

In the FE model, the mobile factor is skilled labour/human capital/entrepreneurs. It captures the empirical evidence, esp. for Europe, according to which skilled workers/entrepreneurial undertakings are typically more mobile than less specialized labour (Forslid and Ottaviano 2003, p. 230). Differently from Forslid and Ottaviano (2003), we envisage a possible situation in which also the entrepreneurs mobility could be limited due to various types of barriers (for example: political, cultural, language and legal barriers).

[2]The building up of a manageable multiregional model, where the number of regions is finite but not limited to a small number, represents one of the research frontiers of the NEG paradigm (see Fujita and Thisse 2009). Recent attempts in this direction are those by Akamatsu and Takayama (2009), Bosker et al. (2010), Ikeda et al. (2010, 2012), Akamatsu et al. (2012).

2.1.1 Basic Assumptions

We consider an economy characterised by R regions (region r goes from 1 to R). There are two sectors, agriculture A and manufacturing M. While there is a unique homogeneous agricultural good, manufacturing involves the production of N differentiated goods or varieties (good i goes from 1 to N). Finally, two types of agents exist, entrepreneurs E and workers L, which are endowed with human capital and labour, respectively. Workers are immobile (but can be reallocated across sectors), whereas entrepreneurs can migrate across regions. Regions can be grouped together into countries/trade blocs C (country c goes from 1 to C).[3] We assume that there is no factor mobility across countries/trade blocs.

2.1.2 Production

The A sector is characterised by perfect competition and constant returns to scale. Production of 1 unit of output requires only workers as input, without loss of generality, we assume that one unit of L is required per unit of otput. The M sector instead is (Dixit–Stiglitz) monopolistically competitive involving increasing returns. It is modelled according to a few basic characteristics: identical firms produce differentiated goods/varieties with the same production technology involving a fixed component, one entrepreneur, and a variable component, unskilled workers, with η units of L required for each unit of the differentiated good. Total cost TC for a firm i corresponds to

$$TC(q_i) = \pi_i + w\eta q_i$$

where q is the output and w the wage rate and where the fixed cost component π_i represents the remuneration of the entrepreneur and the operating profit:

$$\pi_i = p_i q_i - w\eta q_i \tag{1}$$

where p is the price. Given consumers' preference for variety (see below) and increasing returns, each firm will always produce a variety different from those produced by the other firms. Moreover, since one entrepreneur is required for each manufacturing firm, the total number of firms/varieties, N, always equates the total number of entrepreneurs, $E = N$. Denoting by $\lambda_{r,t}$ the share of entrepreneurs located in region r during the time unit t, the number of regional varieties produced in region r during that period is

$$N_{r,t} = \lambda_{r,t} N = \lambda_{r,t} E.$$

[3] Our definition of C does not include only countries but also trade blocs with different degrees of integration. For example, in a model with $C = 2$, one of the two trade areas could represent the EU and the other one any outside country/bloc.

2.1.3 Trade Costs

Distance plays a crucial role in NEG models. Trade between regions can be inhibited by various types of costs that can involve transportation and/or (tariffs or non-tariff) barriers and/or other types of impediments/frictions. In this section, we use a broad definition of trade costs, T. Following Anderson and van Wincoop (2004, pp. 691–692): "[t]rade costs broadly defined, include all costs incurred in getting a good to a final user other than the marginal cost of producing the good itself: transportation costs (both freight costs and time costs), policy barriers (tariffs and non-tariff barriers), information costs, contract enforcement costs, costs associated with the use of different currencies, legal and regulatory costs, and local distribution costs (wholesale and retail)." We represent distance as an exogenous parameter, without specifying its nature and characteristics, which is differentiated across regions. We leave to Sect. 3 a lengthier discussion on the meaning of T in the literature on geographical economics and briefly discuss the case of endogenous trade costs. The general form of the trade cost/distance $(R \times R)$ matrix is

$$\mathbf{T} = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1R} \\ T_{21} & T_{22} & \cdots & T_{2R} \\ \vdots & & \ddots & \vdots \\ T_{R1} & T_{R2} & \cdots & T_{RR} \end{bmatrix}$$

where entry T_{ij} denotes the trade cost (a unidirectional link) between region i and region j. The matrix $\mathbf{T}$ also describes the structure of the transport/trade network between the regions (each region representing a node and the trade cost/distance between two regions the corresponding link). The long-run distribution of the manufacturing activities crucially depends on the shape and properties of such a network structure.

Trade costs are a major component of prices. How trade costs may affect prices crucially depends on the model setting, as we shall see below where, alternatively, trade costs are assumed to be proportional to the [fob/mill] price as in standard NEG models (*iceberg trade costs*) or an addition to the [fob] price as in linear NEG models (*linear trade costs*). For the first type of models, we can apply the standard transformation of iceberg trade costs into trade freeness $\phi_{ij} \equiv T^{1-\sigma}$, with $0 < \phi_{ij} \leq 1$ and with σ to be defined later, obtaining the following $(R \times R)$ matrix

$$\boldsymbol{\phi} = \begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1R} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2R} \\ \vdots & & \ddots & \vdots \\ \phi_{R1} & \phi_{R2} & \cdots & \phi_{RR} \end{bmatrix}$$

When $\phi_{ij} = \phi_{ji}$ $(T_{ij} = T_{ji})$ the above matrices become symmetric. The region is the basic spatial unit. Distance within a region has no impact, which implies that

local trade is costless. This is why, along the main diagonal, we write only 1s for the standard NEG model $T_{ii} = 1$ ($\phi_{ii} = 1$) (for all i) and all 0s for the linear one $T_{ii} = 0$ (for all i).

2.1.4 Consumers' Preferences

Individual (entrepreneur or unskilled worker) preferences are represented by a two-tier utility function, implicitly expressed by

$$U = U(C_M, C_A)$$

The upper-tier concerns the choice between agricultural and manufactured goods, where C_A is the consumption of the agricultural good. The lower-tier concerns the consumption of the manufactured varieties, $C_M = C_M(c_i)$, where c_i represents the consumption of variety i and where $i = 1, \ldots, N$.

The corresponding indirect utility function is:

$$V = U(y, P, p_A)$$

where y is the individual agent (unskilled worker or entrepreneur) income, P is a scalar representing the price composite of the manufactured goods (or price index) and p_A is the price of the agricultural good, which is chosen as the numeraire and set equal to 1 (see below). The properties of $V(\cdot)$ are specified in detail by Pflüger and Südekum (2008): it has the standard properties of continuity and non-satiation. Moreover, it is differentiable in both arguments, with $V_y > 0$ and $V_P < 0$. The two NEG models we consider assume different explicit forms for $U(\cdot)$ (and consequently for $V(\cdot)$).

The individual budget constraint of an individual resident in region r is

$$\sum_{i=1}^{N} p_i c_i + p_A C_A = y + p_A \overline{C}_A \tag{2}$$

where $\overline{C}_A$ is the individual endowment of the agricultural good which is assumed sufficiently large to allow for positive consumption of the numéraire in equilibrium; p_i is the price of variety i inclusive of transport costs.

2.1.5 From the Short to the Long Run

A short-run general equilibrium (SRGE), at time t, is defined for a given spatial distribution of entrepreneurs across the regions, corresponding to the vector of given shares $(\lambda_{1,t}, \lambda_{2,t}, \ldots, \lambda_{R,t})$ and it is instantaneously realised. What characterises a short-run equilibrium is the clearing of all markets: supply equals demand for the

agricultural good and each manufacturer meets the demand for its variety; and by Walras's law simultaneous equilibrium in the product markets implies equilibrium in the regional labour markets. With no transport costs, the price of the agricultural good A is the same across regions and it is set equal to 1, representing the numéraire. It follows that in a SRE, $w = p_A = 1$.

To derive explicitly manufacturing prices, regional price indexes and operating profits, we need an explicit form for the utility function $U(\cdot)$. We anticipate here that, in the standard FE model the price set by the individual manufacturing firm in region r only depends on exogenous parameters, among which the trade costs necessary to deliver commodities in region s, $p = p(T_{rs})$. Instead, in the linear FE model it also depends negatively on the regional share of manufacturing firms $p_{r,t} = p(T_{rs}, \lambda_{r,t})$. Operating profits and price indexes determine real incomes and indirect utilities of entrepreneurs located in region r, $V_{r,t}$.

The shift from a short-run equilibrium t to the following $t+1$ occurs allowing for a change in the share of regional manufacturing activity. Entrepreneurial migration is based on an economic incentive. Specifically, our migration hypothesis involves a discrete time process centred on a comparison between the indirect utility obtained in region r and a weighted average of indirect utilities in all regions—a mechanism resembling the replicator dynamics:

$$\frac{M_{r,t+1} - \lambda_{r,t}}{\lambda_{r,t}} = \gamma_r \left(\frac{V_{r,t} - \sum_{s=1}^{R} \lambda_{s,t} V_{s,t}}{\sum_{s=1}^{R} \lambda_{s,t} V_{s,t}} \right) \tag{3}$$

where $M_{r,t+1}$ denotes the share of entrepreneurs in region r without taking into account any boundary conditions; γ_r is the migration speed, $\gamma_r \geq 0$, and $r = 1, \ldots, R$; when $\gamma_r = 0$, there is no migration. Moreover, the boundary conditions on the shares must hold: $0 \leq \lambda_{r,t} \leq 1$ and $\lambda_{1,t} + \ldots + \lambda_{R,t} = 1$. Expression (3) can be rewritten as:

$$M_{r,t+1} = \lambda_{r,t} \left[1 + \gamma_r \left(\frac{V_{r,t} - \sum_{s=1}^{R} \lambda_{s,t} V_{s,t}}{\sum_{s=1}^{R} \lambda_{s,t} V_{s,t}} \right) \right] = \lambda_{r,t} \left(1 + \gamma_r K_{r,t} \right)$$

In order to model a migration process involving only a subset of regions Z (what we have denoted a country/trade bloc), numbered $z = f, \ldots, l$ with $1 \leq f < l \leq R$ and $Z = l - f + 1 < R$, we write:

$$\frac{M_{z,t+1} - \lambda_{z,t}}{\lambda_{z,t}} = \gamma_z \left(\frac{V_{z,t} - \sum_{m=f}^{l} \lambda_{m,t} V_{m,t}}{\sum_{m=f}^{l} \lambda_{m,t} V_{m,t}} \right) \tag{4}$$

or

$$M_{z,t+1} = \lambda_{z,t} \left[1 + \gamma_z \left(\frac{V_{z,t} - \sum_{m=f}^{l} \lambda_{m,t} V_{m,t}}{\sum_{m=f}^{l} \lambda_{m,t} V_{m,t}} \right) \right] = \lambda_{z,t} \left(1 + \gamma_z K_{z,t} \right)$$

where $0 \leq \lambda_{z,t} \leq \lambda_c$ and $\lambda_{f,t} + \ldots + \lambda_{l,t} = \lambda_c$ and where $0 < \lambda_c \leq 1$ is the (given) share of entrepreneurs located in country c.

A long-run interior equilibrium is defined by a vector $(\lambda_1^*, \ldots, \lambda_R^*)$ such that

$$V_r(\lambda_1^*, \ldots, \lambda_R^*) = \sum_{s=1}^{R} \lambda_s^* V_s(\lambda_1^*, \ldots, \lambda_R^*)$$

for each r. A fully symmetric equilibrium corresponds to $\lambda_1^* = \ldots = \lambda_R^* = R^{-1}$ and a symmetric equilibrium involving only a subset of regions Z to $\lambda_f^* = \ldots = \lambda_l^* = Z^{-1}\lambda_c$.

To evaluate the stability properties of the system of difference equations in (3), we need to compute the following $(R \times R)$ Jacobian matrix:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial M_{1,t+1}}{\partial \lambda_{1,t}} & \frac{\partial M_{2,t+1}}{\partial \lambda_{1,t}} & \cdots & \frac{\partial M_{R,t+1}}{\partial \lambda_{1,t}} \\ \frac{\partial M_{1,t+1}}{\partial \lambda_{2,t}} & \frac{\partial M_{2,t+1}}{\partial \lambda_{2,t}} & \cdots & \frac{\partial M_{R,t+1}}{\partial \lambda_{2,t}} \\ \vdots & & \ddots & \vdots \\ \frac{\partial M_{1,t+1}}{\partial \lambda_{R,t}} & \frac{\partial M_{2,t+1}}{\partial \lambda_{R,t}} & \cdots & \frac{\partial M_{R,t+1}}{\partial \lambda_{R,t}} \end{bmatrix}$$

where

$$\frac{\partial M_{r,t+1}}{\partial \lambda_{s,t}} = 1 + \gamma_r K_{r,t} + \frac{\partial K_{r,t}}{\partial \lambda_{s,t}}$$

$$\frac{\partial K_{r,t}}{\partial \lambda_{s,t}} = \frac{\frac{\partial V_{r,t}}{\partial \lambda_{s,t}}}{\sum_{k=1}^{R} \lambda_{k,t} V_{k,t}} - \left(\frac{2V_{r,t+1} - \sum_{k=1}^{R} V_{k,t} + \sum_{k=1}^{R} \lambda_{k,t} \frac{\partial V_{k,t}}{\partial \lambda_{s,t}}}{\left(\sum_{k=1}^{R} \lambda_{k,t} V_{k,t} \right)^2} \right),$$

with $r = 1, \ldots, R$ and $s = 1, \ldots, R$. When only a subset of regions Z is involved, the matrix $\mathbf{J}$ should be reduced to the following $(Z \times Z)$ submatrix $\mathbf{H}$:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial M_{f,t+1}}{\partial \lambda_{f,t}} & \frac{\partial M_{f+1,t+1}}{\partial \lambda_{f,t}} & \cdots & \frac{\partial M_{l,t+1}}{\partial \lambda_{f,t}} \\ \frac{\partial M_{f,t+1}}{\partial \lambda_{f+1,t}} & \frac{\partial M_{f+1,t+1}}{\partial \lambda_{f+1,t}} & \cdots & \frac{\partial M_{l,t+1}}{\partial \lambda_{f+1,t}} \\ \vdots & & \ddots & \vdots \\ \frac{\partial M_{f,t+1}}{\partial \lambda_{l,t}} & \frac{\partial M_{f+1,t+1}}{\partial \lambda_{l,t}} & \cdots & \frac{\partial M_{l,t+1}}{\partial \lambda_{l,t}} \end{bmatrix}$$

where

$$\frac{\partial M_{z,t+1}}{\partial \lambda_{v,t}} = 1 + \gamma_z K_{z,t} + \frac{\partial K_{z,t}}{\partial \lambda_{v,t}}$$

$$\frac{\partial K_{z,t}}{\partial \lambda_{v,t}} = \frac{\frac{\partial V_{z,t}}{\partial \lambda_{v,t}}}{\sum_{m=f}^{l} \lambda_{m,t} V_{m,t}} - \left(\frac{2V_{z,t+1} - \sum_{m=f}^{l} V_{m,t} + \sum_{m=f}^{l} \lambda_{m,t} \frac{\partial V_{m,t}}{\partial \lambda_{v,t}}}{\left(\sum_{m=f}^{l} \lambda_{m,t} V_{m,t} \right)^2} \right)$$

with $z = f, \ldots, l$ and $v = f, \ldots, l$.

(Local) stability of the system in (3) (or in (4)) requires the eigenvalues of the matrix **J** (or **H**), evaluated at a specific equilibrium, belonging to the interior of the unit circle. We do not pursue this analysis in further detail but we will study two examples below. As we shall see, the existence and stability properties of long-run equilibria depend crucially on the modelling strategy. In Sects. 2.3.2 and 2.4.2, we explore two of those proposed in the literature: the standard FE model and the linear FE model.

Before dealing in detail with these two different modelling strategies, we present in the following subsection an overview of the existing literature on multiregional NEG models.

2.2 Taxonomy of the Literature

As mentioned above, the effect of trade liberalization on the long-term distribution of economic activities is crucially affected by the model specification. The existing contributions to the literature can be grouped into two classes. To the first class of models, where trade liberalization leads to the prevailing of agglomeration forces over dispersion forces, belongs the standard NEG (SN) approach. The analytical framework is typically that introduced by Krugman (1991) and adopted, for example, in the CP and FE model variants [see also the compendia on the NEG approach written by Fujita et al. 1999 (FKV); and Baldwin et al. 2003. A further approach to multi-regional modelling, included into this class, is that put forward by Puga and Venables (1997) (see also Puga and Venables 1996, 1999; and Puga 1999), based on the NEG Vertical Linkages (VL) model developed by Krugman and Venables (1995) and Venables (1996). Here the differentiated goods enter also in the production process as intermediate inputs; firms' demand for intermediates constitutes a further agglomeration force. A third approach put forward by Bosker et al. (2010) can be considered a generalisation of the first two. Indeed, these authors adopts Puga (1999) set-up whose general cost function includes the SN and VL models as special cases (SNVL).

Finally, other authors use the utility function suggested by Pflüger (2004) [Quasi-linear logarithmic (QLL) model] in which the upper-tier is quasi linear with a logarithmic component concerning the N manufactured goods; whereas the lower-tier is still a CES.

In the second class of models, instead, reducing trade costs has the opposite effect with dispersion forces prevailing over agglomeration forces. Here we can distinguish between two subgroups using different analytical structures: (a) that

introduced by Krugman and Elizondo (1996) (KE), in which the typical dispersion force of the standard NEG model represented by the demand from the workers in the A sector, has been substituted by commuting/land costs. Indeed, a consequence of labour migration and firms relocation is the rise in commuting distance and land rentals, reducing the attractiveness of the region for labour migration; (b) that put forward by Ottaviano et al. (2002), OTT, that only differs from the standard NEG approach in the functional form of the utility function and in the specification of trade costs. A direct consequence of this modelling strategy is a CIF price (price at destination) which falls as the number of competing firms rises.

The importance of the different modelling strategies is particularly evident in the case of distance asymmetries (ex. hub and spoke) with respect to an outside country, where a further effect plays a relevant role: the competition effect that originates from trade with the outside region. As noted by Crozet and Koenig Soubeyran (2004), in the context of a standard 2-country 3-region NEG model, when the outside region is sufficiently large firms could agglomerate in the most distant region (spoke) in order to mitigate the competitive pressure originating from the foreign firms. But mostly the SN approach is not able to capture this possibility.

In Table 1, the existing contributions have been systematised on the basis of two main criteria: (a) mobility: on the one side all multiregional models for which there are not, in principle, impediments to factors mobility; on the other multiregional–multicountry models for which mobility is allowed only within a subset of regions (that is, only within a country/trade area) and (b) symmetry: we distinguish on the basis of the symmetric/asymmetric structure of regional sizes and distances. For each work we further identify the type of model (standard NEG, linear NEG, and so on) and the number of regions/countries considered (i.e. the geographical structure: for example, 1-country, R-region, R; 2-country 3-region, $2 \oplus 1$; 3-country 4-region, $1 \oplus 2 \oplus 1$; and so on).

2.3 The Standard FE Model

2.3.1 The General Case

The standard FE model assigns Cobb–Douglas preferences over the choice between consumption of the agricultural good C_A and consumption of the composite of manufactured varieties C_M:

$$U = C_M^{\mu} C_A^{1-\mu}$$

and it assumes a CES utility function to describe the preferences across the manufactured varieties:

$$C_M = \left(\sum_{i=1}^{N} c_i^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}}$$

Table 1 The state of the literature on NEG models

	Spatial mobility					
	Full			Partial		
	Author	Regions	Type	Author	Regions	Type
Symmetric regions						
Full symmetry (a1)	Fujita et al. (1999)	3	SN			
	Castro et al. (2012)	3	SN			
	Gaspar et al. (2013)	3	SN			
	Commendatore and Kubin (2013)	3	SN			
Partial symmetry (a2)	Ikeda et al. (2012)	R	SN	Krugman and Elizondo (1996)	2 ⊕ 1	
	Akamatsu et al. (2012)	R	QLL	Villar (1999)	1 ⊕ 3 ⊕ 1	KE
	Akamatsu and Takayama (2009)	R	SN, QLL	Alonso-Villar (2001)	1 ⊕ 2 ⊕ 1	KE
	Ikeda et al. (2010)	R	SN, QLL	Puga and Venables (1997)	2 ⊕ 1	VL
				Paluzie (2001)	2 ⊕ 1	SN
				Monfort and Nicolini (2000)	2 ⊕ 2	SN
				Monfort and Van Ypersele (2003)	2 ⊕ 2	SN
				Behrens et al. (2007)	2 ⊕ 2	LN
				Behrens (2011)	2 ⊕ 1	LN
				Commendatore et al. (2012)	2 ⊕ 1	SN
				Commendatore et al. (2014)	2 ⊕ 1	SN

(continued)

Table 1 (continued)

	Spatial mobility					
	Full			Partial		
	Author	Regions	Type	Author	Regions	Type
Asymmetric regions						
Asymmetric trade costs (b1)	Krugman (1993)	3	SN, HS	Puga and Venables (1997)	2 ⊕ 1	VL, HS
	Ago et al. (2006)	3	SN, LN, HS	Crozet and Koenig Soubeyran (2004)	2 ⊕ 1	SN, HS
				Brülhart et al. (2004)	2 ⊕ 1	QLL
Full asymmetry (b2)	Forslid (2004)	3	SN	Wang and Zheng (2013b)	2 ⊕ 1	SN, HS
	Baldwin et al. (2003)	3	SN	Wang and Zheng (2013a)	2 ⊕ 1	LN, HS
				Bosker et al. (2010)		R, SNVL

Codes for the structure of regions as follows: 3 = 1-country, 3-region world. *R* =1-country, R-region world. 2 ⊕ 1 =2-country, 3-region world. 2 ⊕ 2 =2-country, 4-region world. 1 ⊕ 2 ⊕ 1 =3-country, 4-region world. 1 ⊕ 3 ⊕ 1 =3 country, 5-region world. Codes for the type of model: *SN* standard NEG, *VL* NEG Vertical Linkages, *QLL* quasi-logarithmic, *KE* Krugman and Elizondo (1996) type, *HS* Hub and Spoke, *LN* Linear NEG, *R* One country with more than three regions, *SNVL* Standard NEG Vertical Linkages

A. Factors of production can freely move across all the regions
B. Mobility is allowed only within subsets/subgroups of regions (typical examples: countries, free-trade areas, and similar)

(a) The regions are symmetric:

1. Full symmetry: all regions are symmetric (both in sizes and distances)
2. A subset of regions is symmetric (i.e., all the regions belonging to the same country/free-trade area are characterised by the same transport costs with the outside world; or the regions are located on the vertices of a symmetric geometric figure with a number of sides larger than 3)

(b) The regions are asymmetric:

1. Asymmetric distances (even within countries there are differences in transport costs) hub and spoke (HS) is a special case according to which the spoke is linked to the outside world only through the hub (or the route through the hub is the cheapest one). Instead, a more general "hub effect" (see Krugman (1993)) emerges when a region is more accessible compared to neighbouring regions
2. Asymmetric distances and sizes

where c_i represents the quantity consumed of the variety i, with $i = 1, \ldots, N$; σ the constant elasticity of substitution/taste for variety: the closer σ to 1, the greater is consumer's taste for variety, with $\sigma > 1$; and μ and $1 - \mu$ represent the income shares devoted to the manufactured varieties and to the homogeneous agricultural good, respectively, with $0 < \mu < 1$.

Following profit maximization, each firm sets for its variety a (mill) price $\tilde{p}$ on the basis of a perceived elasticity of $-\sigma$ (under the assumption of symmetric firm behaviour, we drop the subscript i):

$$\tilde{p} = \frac{\sigma}{\sigma - 1}\eta$$

Considering the short-run equilibrium, we fix for period t the regional shares of entrepreneurs and add the time subscript.

The overall demand for each variety (where it is also taken into account the part that is melting along the way because of iceberg costs) corresponds to:

$$d_{r,t} = \left(\sum_{s=1}^{R} \mu Y_{s,t} P_{s,t}^{\sigma-1} T_{rs}^{1-\sigma}\right) \tilde{p}^{-\sigma} = \left(\sum_{s=1}^{R} \frac{s_{s,t}}{\Delta_{s,t}} \phi_{rs}\right) \tilde{p}^{-1} \frac{\mu Y}{E}$$

where

$$P_{r,t} = \left(\sum_{s=1}^{R} n_{s,t}^{1-\sigma} \tilde{p}^{1-\sigma} T_{rs}^{1-\sigma}\right)^{\frac{1}{1-\sigma}} = \Delta_{r,t}^{\frac{1}{1-\sigma}} E^{\frac{1}{1-\sigma}} \tilde{p}$$

is the price index facing consumers in region r; $Y_{s,t}$ represents income and expenditure in region s; $s_{s,t} = \frac{Y_{s,t}}{Y}$ denotes region s's share in expenditure in total (world) income Y and $s = 1, \ldots, R$. Moreover, we have defined

$$\Delta_{r,t} = \lambda_{1,t}\phi_{r1} + \lambda_{2,t}\phi_{r2} + \lambda_{3,t}\phi_{r3} + \ldots + \lambda_{R,t}\phi_{rR} = \sum_{s=1}^{R} \lambda_{s,t}\phi_{rs}.$$

The operating profit in region r corresponds to

$$\pi_{r,t} = \left(\sum_{s=1}^{R} \mu Y_{s,t} P_{s,t}^{\sigma-1} T_{rs}^{1-\sigma}\right) \frac{p^{1-\sigma}}{\sigma} = \left(\sum_{s=1}^{R} \frac{s_{s,t}}{\Delta_{s,t}} \phi_{rs}\right) \frac{\mu Y}{\sigma E} \tag{5}$$

Thus, the indirect utility of the entrepreneur located in region r can be expressed as:

$$V_{r,t} = \frac{\pi_{r,t}}{P_{r,t}^{\mu}}$$

In order to study the time evolution of the system as well as the existence and (local) stability properties of the long-run equilibria, this expression should be inserted into the migration hypothesis (3) (or alternatively into (4)). Then the analysis should proceed by evaluating the corresponding Jacobian matrix **J** (or **H**), verifying for which parameter combinations the eigenvalues condition is satisfied. Given the complexity of the system and the objectives of this review, we limit ourselves to the exploration of the special case of a $2 \oplus 1$ economy. That is, an economy composed of two countries/trade blocs and three regions. We turn to such analysis in the following subsection.

2.3.2 The 3-Region Model

We explore a special case of a 3-region, 2 country model, representing a $2 \oplus 1$ economy as defined above, where entrepreneurs can move only between region 1 and region 2. The two regions represent a customs union with symmetric trade costs within the union, denoted by T_S; trade costs with respect to region 3 (rest of the world) is T_L, with $T_S < T_L$. Moreover, Region 1 and 2 have identical sizes. The trade costs and trade freeness matrices become:

$$\mathbf{T} = \begin{bmatrix} 1 & T_S & T_L \\ T_S & 1 & T_L \\ T_L & T_L & 1 \end{bmatrix}; \boldsymbol{\phi} = \begin{bmatrix} 1 & \phi_S & \phi_L \\ \phi_S & 1 & \phi_L \\ \phi_L & \phi_L & 1 \end{bmatrix}$$

That is, in terms of trade flows, the economy can be represented as a small network composed of three nodes (located at the vertices of an acute isosceles triangle) and three links (the tree sides of the triangle, with the link between region 1 and 2 representing the shortest one).

We denote with $\tilde{E}$ the number of entrepreneurs which are free to move between region 1 and 2 (the customs union), with $\tilde{E} = N_{1,t} + N_{2,t}$; consequently $\overline{E} = E - \tilde{E} = \overline{N}_3$ represents the number of immobile entrepreneurs located in region 3. Moreover, we denote by x_t and $1 - x_t$ the shares of mobile entrepreneurs located in region 1 and 2, respectively.

We have that:

$$E = \tilde{E} + \overline{E} = N_{1,t} + N_{2,t} + \overline{N}_3 = \lambda_{1,t} E + \lambda_{2,t} E + \overline{\lambda}_3 E = x_t \tilde{E} + (1 - x_t)\tilde{E} + \overline{E}.$$

Finally, after defining $\frac{\tilde{E}}{E} = \tilde{n}$, we can write

$$\lambda_{1,t} = x_t \frac{\tilde{E}}{E} = x_t \tilde{n}; \lambda_{2,t} = (1 - x_t)\frac{\tilde{E}}{E} = (1 - x_t)\tilde{n}; \lambda_{3,t} = \overline{\lambda}_3 = 1 - \tilde{n}.$$

Taking into account these expressions and that we are considering a 3-region economy, the regional price indexes can be expressed as:

$$P_{r,t} = \left(N_{1,t}^{1-\sigma}\tilde{p}^{1-\sigma}T_{r1}^{1-\sigma} + N_{2,t}^{1-\sigma}\tilde{p}^{1-\sigma}T_{r2}^{1-\sigma} + \overline{N}_3\tilde{p}^{1-\sigma}T_{r3}^{1-\sigma}\right)^{\frac{1}{1-\sigma}}$$
$$= \Delta_{r,t}^{\frac{1}{1-\sigma}}\left(\frac{\sigma}{\sigma-1}\eta\right)E^{\frac{1}{1-\sigma}}$$

where $\Delta_{r,t} = x_t\tilde{n}\phi_{r1} + (1-x_t)\tilde{n}\phi_{r2} + (1-\tilde{n})\phi_{r3}$.

The operating profit in region r is

$$\pi_{r,t} = \left(s_{1,t}\Delta_{1,t}^{-1}\phi_{r1} + s_{2,t}\Delta_{2,t}^{-1}\phi_{r2} + s_{3,t}\Delta_{3,t}^{-1}\phi_{r3}\right)\frac{\mu Y}{\sigma E}$$

Regional incomes/expenditures are:

$$Y_{r,t} = L_r + \lambda_{r,t}\pi_{r,t}E.$$

Taking into account that $\phi_{ii} = 1$, $\phi_{ij} = \phi_{ji}$, $\phi_{12} = \phi_S$, $\phi_{13} = \phi_{23} = \phi_L$; and that region 1 and 2 have identical proportions of unskilled workforce (and therefore identical wage shares): $\frac{L_1}{L} = \frac{L_2}{L} = \theta$, which implies $\frac{L_3}{L} = 1 - 2\theta$, the regional income shares s_r for the case $R = 3$ take the form:

$$s_{1,t} = \frac{(\sigma-\mu)\theta + \mu\tilde{n}x_t\left[\frac{\phi_L}{\Delta_{3,t}} - \frac{\left(\frac{\phi_L}{\Delta_{3,t}} - \frac{\phi_S}{\Delta_{2,t}}\right)\left((\sigma-\mu)\theta + \frac{\mu\tilde{n}(1-x_t)\phi_L}{\Delta_{3,t}}\right)}{\sigma - \mu\tilde{n}(1-x_t)\left(\frac{1}{\Delta_{2,t}} - \frac{\phi_L}{\Delta_{3,t}}\right)}\right]}{\left[\sigma - \mu\tilde{n}x_t\left(\frac{1}{\Delta_{1,t}} - \frac{\phi_L}{\Delta_{3,t}}\right)\right]\left[1 - \frac{\left(\frac{\phi_L}{\Delta_{3,t}} - \frac{\phi_S}{\Delta_{2,t}}\right)\left(\frac{\phi_L}{\Delta_{3,t}} - \frac{\phi_S}{\Delta_{1,t}}\right)\mu^2\tilde{n}^2x_t(1-x_t)}{\left[\sigma - \mu\tilde{n}(1-x_t)\left(\frac{1}{\Delta_{2,t}} - \frac{\phi_L}{\Delta_{3,t}}\right)\right]\left[\sigma - \mu\tilde{n}x_t\left(\frac{1}{\Delta_{1,t}} - \frac{\phi_L}{\Delta_{3,t}}\right)\right]}\right]}$$

$$s_{2,t} = \frac{(\sigma-\mu)\theta + \mu\tilde{n}(1-x_t)\left[\frac{\phi_L}{\Delta_{3,t}} - \frac{\left(\frac{\phi_L}{\Delta_{3,t}} - \frac{\phi_S}{\Delta_{1,t}}\right)\left((\sigma-\mu)\theta + \frac{\mu\tilde{n}x_t\phi_L}{\Delta_{3,t}}\right)}{\sigma - \mu\tilde{n}x_t\left(\frac{1}{\Delta_{1,t}} - \frac{\phi_L}{\Delta_{3,t}}\right)}\right]}{\left[\sigma - \mu\tilde{n}(1-x_t)\left(\frac{1}{\Delta_{2,t}} - \frac{\phi_L}{\Delta_{3,t}}\right)\right]\left[1 - \frac{\left(\frac{\phi_L}{\Delta_{3,t}} - \frac{\phi_S}{\Delta_{2,t}}\right)\left(\frac{\phi_L}{\Delta_{3,t}} - \frac{\phi_S}{\Delta_{1,t}}\right)\mu^2\tilde{n}^2x_t(1-x_t)}{\left[\sigma - \mu\tilde{n}(1-x_t)\left(\frac{1}{\Delta_{2,t}} - \frac{\phi_L}{\Delta_{3,t}}\right)\right]\left[\sigma - \mu\tilde{n}x_t\left(\frac{1}{\Delta_{1,t}} - \frac{\phi_L}{\Delta_{3,t}}\right)\right]}\right]}$$

$$s_{3,t} = 1 - s_{1,t} - s_{2,t}.$$

Concerning the central dynamic equation, given that the share of entrepreneurs located in region 3 is given, $\overline{\lambda}_3 = 1 - \tilde{n}$, the migration law only involves regions 1 and 2. Therefore, the relevant state variable is x_t. We can write

$$\frac{Z_{t+1} - x_t}{x_t} = \gamma\left((1-x_t)\frac{V_{1,t} - V_{2,t}}{x_tV_{1,t} + (1-x_t)V_{2,t}}\right)$$

where $Z_{t+1} = \frac{M_{t+1}}{\tilde{n}} = Z(x_t)$ and where the indirect utility functions correspond to $V_{1,t} = \pi_{1,t}P_{1,t}^{-\mu} = V_1(x_t)$ and $V_{2,t} = \pi_{2,t}P_{2,t}^{-\mu} = V_2(x_t)$. Notice also that γ represents the entrepreneurial migration speed, which is positive only for movements between region 1 and 2 and it is equal to zero otherwise. After simple

manipulations:

$$\frac{Z_{t+1}-x_t}{x_t}=\gamma\left((1-x_t)\frac{T(x_t)}{1+x_tT(x_t)}\right)$$

with:

$$T(x_t)=\frac{V_{1,t}}{V_{2,t}}-1=\frac{\frac{s_{1,t}}{\Delta_{1,t}}+\frac{s_{2,t}}{\Delta_{2,t}}\phi_S+\frac{1-s_{1,t}-s_{2,t}}{\Delta_{3,t}}\phi_L}{\frac{s_{1,t}}{\Delta_{1,t}}\phi_S+\frac{s_{2,t}}{\Delta_{2,t}}+\frac{1-s_{1,t}-s_{2,t}}{\Delta_{3,t}}\phi_L}\left(\frac{\Delta_{2,t}}{\Delta_{1,t}}\right)^{\frac{\mu}{1-\sigma}}-1.$$

Taking into account the constraints, $0\leq x_t\leq 1$, we have the full dynamical model:

$$x_{t+1}=\begin{cases}0 & if \quad Z(x_t)<0\\ Z(x_t) & if\ 0\leq Z(x_t)\leq 1\\ 1 & if \quad Z(x_t)>1.\end{cases}$$

In Commendatore et al. (2014) it is developed the full dynamical analysis for this map. The less ambitious objective here is to compare the properties of the two models, the standard and the linear NEG model, in correspondence of the so-called 'break' point of the trade freeness parameter ϕ_S. The break point of trade freeness (or, through a transformation, of trade costs) is that value ϕ_S^b that satisfies the condition:

$$Z'\left(\frac{1}{2}\right)=1 \quad\Longleftrightarrow\quad T'\left(\frac{1}{2}\right)=0.$$

For $0<\tilde{n}<1$, ϕ_S^b corresponds to the positive root of a quadratic equation whose expression is quite long and complicated. From simulations it appears that for some meaningful parameter combinations, but not for all, there exists a ϕ_S^b belonging to the interval (0,1).

In Commendatore et al. (2012), the authors making the simplifying assumption of absent industrial sector in region 3, $\tilde{n}=1$, are able to obtain the following relatively simple expression:

$$\phi_S^b=\frac{(\sigma-\mu)[2\theta(\sigma-1)-\mu]}{2\theta(\sigma-\mu)(\sigma-\mu 1)+\mu(3\sigma-2+\mu)}<1$$

Notice that, when $\tilde{n}=1$, ϕ_L has no effect on ϕ_S^b. That is, the distance between the union and the outside region has no impact on the stability properties of the symmetric equilibrium and on the location pattern.

Following Pflüger and Südekum (2008), the local stability properties of the symmetric equilibrium and, therefore, the specific location pattern depend on the characteristics of the bifurcation value. The typical bifurcation scenario of a standard 2-region FE model is catastrophic agglomeration corresponding to a

'sub-critical' pitchfork bifurcation. However, a smoother agglomeration process can emerge in correspondence of a supercritical pitchfork bifurcation with the emergence of two (locally) stable interior equilibria.

From the theory of dynamical systems (see Wiggins 2003; see also Pflüger and Südekum 2008), in correspondence of a pitchfork bifurcation, that is, when $\phi_S = \phi_S^b$ and $x^* = \frac{1}{2}$, the following condition must hold: $\frac{\partial^2 Z(x_t)}{\partial x_t^2} = 0$ corresponding to

$$\frac{\partial^2 T\,(x_t)}{\partial x_t^2} = 0 \Leftrightarrow \frac{\partial^2 \left(\frac{V_1(x_t) - V_2(x_t)}{V_2(x_t)}\right)}{\partial x_t^2} = 0$$

It can be shown that due to the symmetry of the map $Z(x_t)$ this condition is always satisfied.[4]

The pitchfork bifurcation is sub-critical or supercritical depending on the sign (positive or negative, respectively) of the following third order derivative, evaluated at $\phi_S = \phi_S^b$ and $x^* = \frac{1}{2}$:

$$\left.\frac{\partial^3 Z(x_t)}{\partial x_t^3}\right|_{x^*=\frac{1}{2},\ \phi_S=\phi_S^b} = \left.\frac{\partial^3 T(x_t)}{\partial x_t^3}\right|_{x^*=\frac{1}{2},\ \phi_S=\phi_S^b}$$

As shown in Commendatore et al. (2014), when $\tilde{n} = 1$, the sign of this derivative depends on the sign of the following expression:

$$\begin{aligned}\Xi \equiv\ & 12(\sigma-1)^2(\sigma-\mu)\theta^2 + [2(\sigma-3)\mu^2 + 4(\sigma-1)^2(3\mu-\sigma)]\theta \\ & - \mu(\mu+\sigma-1)[\mu + 2(\sigma-1)\end{aligned}$$

Commendatore et al. (2012) prove the existence of a value of $\theta \equiv \tilde{\theta}$ lying in the interval (0,1) such that $\Xi < 0$ for $\theta < \tilde{\theta}$, determining a supercritical pitchfork bifurcation, and $\Xi > 0$ for $\theta > \tilde{\theta}$, corresponding to a subcritical one. That is, the size of the third region matters determining the location pattern at the bifurcation point: which is smooth when the size of the outside region is sufficiently large and it is catastrophic when the union becomes sufficiently large in relative terms. We envisage that, by continuity, a similar result must hold also when we allow for a local manufacturing sector in region 3 (see Commendatore et al. 2014). Indeed, these authors confirm via simulations for the more general case, $0 < \tilde{n} < 1$, the possible emergence of the two different bifurcation scenarios. Moreover, it is also possible to show via simulations that the break point is decreasing in $\tilde{n}$ and in ϕ_L and it is increasing in θ. That is, increasing competition from outside counterbalance

[4] Indeed, at the bifurcation point, this condition can be reduced to $\frac{1}{V_2(x_t)}\left(\frac{\partial^2 V_1(x_t)}{\partial x_t^2} - \frac{\partial^2 V_2(x_t)}{\partial x_t^2}\right) = 0$, the validity of which can be shown using the same procedure adopted by Pflüger and Südekum (2008, p. 46).

the agglomeration forces; whereas increasing internal and/or external demand reinforces them.

2.4 The Linear FE Model

2.4.1 The General Case

The linear FE model adopts a quasi-linear utility function composed of two parts: a quadratic function defining the preferences across the M goods and a linear component for the consumption of the A good:

$$U = \alpha \sum_{i=1}^{N} c_i - \left(\frac{\beta - \delta}{2}\right) \sum_{i=1}^{N} c_i^2 - \frac{\delta}{2} \left(\sum_{i=1}^{N} c_i\right)^2 + C_A \tag{6}$$

where α represents the intensity of preferences for the manufactured varieties, with $\alpha > 0$; δ represents the degree of substitutability across those varieties, with $\delta > 0$; and where the taste for variety is measured by the (positive) difference $\beta - \delta > 0$.

Solving for C_A the budget constraint (2), substituting into the utility function (6) and then differentiating with respect to c_i, we obtain the following first order conditions ($i = 1, \ldots, N$):

$$\frac{\partial U}{\partial c_i} = \alpha - (\beta - \delta)\, c_i - \delta \sum_{i=1}^{N} c_i - p_i = 0$$

from which

$$p_i = \alpha - (\beta - \delta)\, c_i - \delta \sum_{i=1}^{N} c_i .$$

The linear demand function is

$$\begin{aligned} c_i(p_1, \ldots, p_N) &= \frac{\alpha}{(N-1)\delta + \beta} - \frac{1}{\beta - \delta} p_i + \frac{\delta}{(\beta - \delta)[(N-1)\delta + \beta]} \sum_{i=1}^{N} p_i \\ &= a - (b + Nc)\, p_i + cP \end{aligned}$$

where $P = \sum_{i=1}^{N} p_i$ and where $p_i \leq p^{\max} \equiv \frac{a + cP}{b + cN}$. The indirect utility is given by:

$$V = S + y + \overline{C}_A$$

where S corresponds to the consumer's surplus :

$$S = U(c_i(p_i),\ i \in [0,n]) - \sum_{i=1}^{N} p_i c_i(p_i) - C_A$$

$$= \frac{a^2 N}{2b} + \frac{b + cN}{2} \sum_{i=1}^{N} p_i^2 - aP - \frac{c}{2} P^2. \tag{7}$$

Moving to the short-run equilibrium, we fix for period t the regional shares of entrepreneurs.

The demand faced by a representative firm in each region is:

$$c_{rs,t} = a - (b + cN)\, p_{rs,t} + c P_{s,t} \tag{8}$$

where $c_{rs,t}$ is the demand of a resident in region s (with $s = 1, \ldots, R$) for a good produced in region r (with $r = 1, \ldots, R$); $p_{rs,t}$ is the price of a variety produced in region r and consumed in region s and P_s is the price index in region s, with

$$P_{s,t} = \lambda_{1,t} E p_{1s,t} + \lambda_{2,t} E p_{2s,t} + \ldots + \lambda_{R,t} E p_{Rs,t.} = \sum_{k=1}^{R} \lambda_{k,t} E p_{ks,t}.$$

Notice that, following from the assumption of symmetric behaviour of firms, prices differ across region—segmenting markets—only because of transport costs.

Short-run equilibrium requires that in each segmented market demand equals supply:

$$c_{rs,t} = q_{rs,t} \tag{9}$$

where $q_{rs,t}$ is the output produced in region r that is brought to a market in region s.

The operating profit of a representative firm in region r is:

$$\pi_{r,t} = \sum_{s=1}^{R} (p_{rs,t} - \eta - T_{rs}) q_{rs,t}\, (L_s + \lambda_{s,t} E) \tag{10}$$

Note that, given the purposes of this work, we assume that all regions trade with each other; this corresponds to the second case considered by Behrens (2011) of well-connected regions.

From the profit maximization procedure and market segmentation, considering further that $N = E$, the first order conditions follow:

$$\frac{\partial \pi_{r,t}}{\partial p_{rs,t}} = [a + (\eta + T_{rs})(b + cE) + cP_{s,t} - 2p_{rs,t}(b + cE)]\, (L_s + \lambda_{s,t} E) = 0$$

and therefore

$$p_{rs,t} = \frac{a + cP_{s,t} + (\eta + T_{rs})(b + cE)}{2(b + cE)} = \frac{p_{s,t}^{\max}}{2} + \frac{\eta + T_{rs}}{2} \tag{11}$$

with $r = 1, \ldots, R$; $s = 1, \ldots, R$ and $p_{s,t}^{\max} = \frac{a + cP_{s,t}}{b + cE}$.

The price index is:

$$P_{s,t} = \frac{a + (b + cE)\sum_{k=1}^{R} \lambda_{k,t}(\eta + T_{ks})}{2b + cE} E. \tag{12}$$

From (11) and (12), we obtain

$$p_{rs,t} = \frac{2[a + \eta(b + cE)] + cE\sum_{k=1}^{R} \lambda_{k,t} T_{ks}}{2(2b + cE)}. \tag{13}$$

According to (13), prices set by firms depend not only on local demand but also on the distribution of firms across space (extending the result of Ottaviano et al. 2002, to the R-region case; see also the discussion below).

We assume that all regions trade with each other. In order to have trade across all regions $\frac{p^{\max} + T_{sr}}{2} > T_{sr}$ for all s and r, therefore $T_{sr} < p^{\max}$ should always hold.

Taking into account (8)–(10), (12) and (13), it can be easily deduced that the operating profit is a function of entrepreneurial shares:

$$\pi_{r,t} = \pi(\lambda_{1,t}, \ldots, \lambda_{R,t}).$$

Similarly, from (7), (12) and (13), taking into account that $N = E$, the surplus of the individual entrepreneur can be expressed as:

$$S_{r,t} = S(\lambda_{1,t}, \ldots, \lambda_{R,t}).$$

Finally, inserting operating profits in the expression for the indirect utility, it is possible to obtain:

$$V_{r,t} = S_{r,t} + \pi_{r,t} y + \overline{C}_A = V_r(\lambda_{1,t}, \ldots, \lambda_{R,t}).$$

As we did before for the standard FE model, in order to study the long-run properties of the linear FE, we proceed considering, in the following subsection, the simple case of a $2 \oplus 1$ economy, that is, an economy composed of two countries/trade blocs and three regions.

2.4.2 The 3-Region Model

As before we consider a $2 \oplus 1$ economy. Our analysis is similar to that developed by Behrens (2011) for the case of a linear CP model. For the linear 3-region model, representing a $2 \oplus 1$ economy, the trade costs matrix corresponds to

$$\mathbf{T} = \begin{bmatrix} 0 & T_S & T_L \\ T_S & 0 & T_L \\ T_L & T_L & 0 \end{bmatrix}$$

that describes a small trade network composed of two nodes close to each other and located at a symmetric distance from a farer third node (see above).

Recalling the above notation, $\tilde{E}$ and $\overline{E}$ are the number of entrepreneurs that can move between regions 1 and 2 and those that are immobile and located in region 3, respectively; x_t and $1 - x_t$, are the corresponding shares of mobile entrepreneurs located in regions 1 and 2; and finally $\frac{\tilde{E}}{E} = \tilde{n}$. It follows that

$$\begin{aligned} \lambda_{1,t} &= x_t \frac{\tilde{E}}{E} = x_t \tilde{n} \\ \lambda_{2,t} &= (1 - x_t) \frac{\tilde{E}}{E} = (1 - x_t) \tilde{n} \\ \lambda_{3,t} &= \overline{\lambda}_3 = 1 - \tilde{n}. \end{aligned}$$

Taking into account these expressions and that we are considering a $2 \oplus 1$ economy, the prices set by firms in the three regional markets are:

$$p_{rs,t} = \frac{a + cP_{s,t} + (\eta + T_{rs})(b + cE)}{2(b + cE)} = \frac{p_{s,t}^{\max}}{2} + \frac{\eta + T_{rs}}{2} \tag{14}$$

with $r = 1, \ldots, 3$; $s = 1, \ldots, 3$ and $p_{s,t}^{\max} = \frac{a + cP_{s,t}}{b + cE}$ and the price indexes are:

$$P_{s,t} = \frac{a + (b + cE) \sum_{k=1}^{3} \lambda_{k,t} (\eta + T_{ks})}{2b + cE} E. \tag{15}$$

It is interesting to note that, due to the symmetric distance from region 1 and 2, the price index in region 3 does not depend on the spatial distribution of economic activities in those regions. From (14) and (15) the equilibrium prices are:

$$p_{rs,t} = \frac{2[a + \eta(b + cE)] + cE \sum_{k=1}^{3} \lambda_{k,t} T_{ks}}{2(2b + cE)}. \tag{16}$$

Taking into account our symmetry assumption according to which θ is the equal share of unskilled labour in region 1 and 2, the equilibrium short-run profits in the

segmented markets for regions 1 and 2 are:

$$\begin{aligned}
\pi_{11,t} &= (p_{11,t} - \eta)^2 (b + cE)\,(\theta L + x_t \tilde{n} E) \\
\pi_{12,t} &= (p_{12,t} - \eta - T_S)^2 (b + cE)[\theta L + (1 - x_t)\tilde{n} E] \\
\pi_{13,t} &= (p_{13,t} - \eta - T_L)^2 (b + cE)[(1 - 2\theta) L + (1 - \tilde{n}) E] \\
\pi_{21,t} &= (p_{21,t} - \eta - T_S)^2 (b + cE)\,(\theta L + x_t \tilde{n} E) \\
\pi_{22,t} &= (p_{22,t} - \eta)^2 (b + cE)[\theta L + (1 - x_t)\tilde{n} E] \\
\pi_{23,t} &= (p_{23,t} - \eta - T_L)^2 (b + cE)[(1 - 2\theta) L + (1 - \tilde{n}) E].
\end{aligned}$$

Note that by symmetry $\pi_{13,t} = \pi_{23,t}$. The consumer surpluses in region 1 and 2 are:

$$S_{1,t} = \frac{a^2 E}{2b} + \frac{b + cE}{2} \sum_{k=1}^{3} \lambda_{k,t} E p_{k1,t}^2 - a P_{1,t} - \frac{c}{2} P_{1,t}^2$$

$$S_{2,t} = \frac{a^2 E}{2b} + \frac{b + cE}{2} \sum_{k=1}^{3} \lambda_{k,t} E p_{k2,t}^2 - a P_{2,t} - \frac{c}{2} P_{1,t}^2$$

Finally, the indirect utilities for the entrepreneurs located in region 1 and 2 are

$$\begin{aligned}
V_{1,t} &= S_{1,t} + \pi_{1,t} + \overline{C}_A \\
V_{2,t} &= S_{2,t} + \pi_{2,t} + \overline{C}_A
\end{aligned} \tag{17}$$

Moving to the analysis of the long run, the evolution of the system—and, specifically, the change through time of the state variable x_t, with $0 \leq x_t \leq 1$—is governed by the comparison of indirect utilities between region 1 and 2, encapsulated in the full dynamical system

$$x_{t+1} = \begin{cases} 0 & if \quad K(x_t) < 0 \\ K(x_t) & if \;\; 0 \leq K(x_t) \leq 1 \\ 1 & if \quad K(x_t) > 1 \end{cases}$$

where

$$\begin{aligned}
K(x_t) &= x_t \left[1 + \gamma \left((1 - x_t) \frac{V_{1,t} - V_{2,t}}{x_t V_{1,t} + (1 - x_t) V_{2,t}} \right) \right] \\
&= x_t \left[1 + \gamma \left((1 - x_t) \frac{H(x_t)}{1 + x_t H(x_t)} \right) \right] \\
H(x_t) &= \frac{V_{1,t}}{V_{2,t}} - 1
\end{aligned}$$

and where $V_{1,t} = V_1(x_t)$ and $V_{2,t} = V_2(x_t)$ correspond to the indirect utilities in (17).[5] The break point satisfies the following condition:

$$\left.\frac{\partial K(x_t)}{\partial x_t}\right|_{x_t=\frac{1}{2}} = 1 \iff \left.\frac{\partial H(x_t)}{\partial x_t}\right|_{x_t=\frac{1}{2}} = 0.$$

The last expression can be rewritten as

$$\left.\frac{\partial H(x_t)}{\partial x_t}\right|_{x_t=\frac{1}{2}} = \left.\frac{\frac{\partial V_{1,t}}{\partial x_t} - \frac{\partial V_{2,t}}{\partial x_t}}{V_{1,t}}\right|_{x_t=\frac{1}{2}} = 0$$

where

$$\frac{\partial V_{1,t}}{\partial x_t} - \frac{\partial V_{2,t}}{\partial x_t} = \frac{\tilde{n}E(b+cE)T_S}{4(2b+cE)^2}\left(A(T_S, T_L, \theta, \tilde{n})\right).$$

The term $A(\cdot)$ is increasing in T_L and decreasing in T_S. Indeed:

$$\frac{\partial A(\cdot)}{\partial T_L} = 2(1-\tilde{n})(4b+3cE)E > 0$$
$$\frac{\partial A(\cdot)}{\partial T_S} = -\left[12b(b+cE) + c^2E^2(3-\tilde{n}) + 4c\theta L(2b+cE)\right] < 0.$$

This extends the results of Behrens (2011) according to which reducing trade costs between region 1 and 2 favours agglomeration inside the Union, in line with the standard NEG prediction; whereas reducing distance with respect to the outside world may favour the dispersion of manufacturing activities. Moreover, $A(\cdot)$ is decreasing in θ and $\tilde{n}$:

$$\frac{\partial A(\cdot)}{\partial \theta} = -4cL(2b+cE)T_S < 0$$
$$\frac{\partial A(\cdot)}{\partial \tilde{n}} = -cE\left[(8bL+6cE)T_L - cET_S\right] < 0$$

since $T_L > T_S$. An increase in the size of the foreign demand and an increase in foreign competition have a destabilising effect on the symmetric equilibrium as in the standard FE model.

Given the much weaker nonlinearity of the central map $K(x_t)$ compared to $Z(x_t)$, we can also easily compute the break point for the transport cost:

[5] We leave the full dynamical analysis of the system to future work.

$$T_S^b = \frac{2cE(1-\tilde{n})(4b+3cE)T_L + 8(3b+2cE)(a-b\eta)}{4\theta Lc(2b+cE) + [c^2(3-\tilde{n})E^2 + 12b(b+cE)}$$

which is certainly positive for $\eta < \frac{a}{b}$. Concerning the nature of the bifurcation point, we have that at the symmetric equilibrium (and for any T_S)

$$\frac{\partial^2 H(x_t)}{\partial x_t^2} = 0 \Rightarrow \frac{\partial^2 \left(\frac{V_1(x_t)-V_2(x_t)}{V_2(x_t)}\right)}{\partial x_t^2} = 0.$$

Moreover,

$$\left.\frac{\partial^3 K(x_t)}{\partial x_t^3}\right|_{x^*=\frac{1}{2},\ T_S^b} = \left.\frac{\partial^3 H(x_t)}{\partial x_t^3}\right|_{x^*=\frac{1}{2},\ \phi_S=\phi_S^b} = 0.$$

This is due to the fact that $V_1(x_t)$ and $V_1(x_t)$ are quadratic functions and therefore

$$\frac{\partial^3 V_1(x_t)}{\partial x_t^3} = \frac{\partial^3 V_2(x_t)}{\partial x_t^3}.$$

The specific location pattern of the linear FE $2\oplus 1$ model is characterised by an atypical pitchfork bifurcation with an immediate jump from a symmetric equilibrium to a CP equilibrium as soon as the break point is crossed from right to left ($T_S < T_S^b$).

In summary, with respect to the case of a 3-region 2-country economy, the most important difference between the standard and the linear FE model is the effect of "external" trade liberalization: in the linear FE model, with a linear demand function, prices are negatively affected by the lowering of trade costs. Thus, the competition effect becomes stronger by reducing T_L; whereas in the standard FE model with an isoelastic demand function, the competition effect is weaker and agglomeration prevails both with "internal" ("domestic") and "external" ("international") trade liberalization.

This analysis has shown that trade costs play a crucial role in a multiregional setting. However, in most of the NEG literature (mainly concerned with a 2-region framework) very simple and not too realistic assumptions are adopted to describe trade barriers (e.g., they are uniform across space, exogenous, indifferentiated, and so on). In the following section, we review that part of the NEG theoretical and empirical literature that addresses this issue. These studies provide necessary ingredients to be added in future work to the description of geographical and non-geographical distance represented by the trade matrices presented above.

3 Some Considerations on More Sophisticated Definitions of Transport Costs

Trade costs is an omnibus term for all the costs incurred by a commodity from factory door to final retail: except for the marginal cost of production, all prices observed in the real world include trade costs reflecting the transformation of goods across time and space. These costs arise from a number of distinct sources: (a) transportation costs (both freight costs and time costs); (b) policy barriers (tariff and nontariff barriers); (c) information costs; (d) contract enforcement costs; (e) costs arising from the use of different currencies; (f) legal and regulatory costs; (g) local distribution costs (wholesale and retail). While some components of trade costs can be easily computed, given that the corresponding figures are widely available, others are intrinsically opaque: because of this lack of availability and objectivity, trade theory is particularly helpful in setting the econometric framework to overcome the issue of missing data. Using an augmented gravity model with various types of international data, Anderson and van Wincoop (2004) are able to quantify the extent of various sources of costs. A rough estimate of the tax equivalent of "representative" trade costs for industrialized countries is. This number breaks down as due to transportation costs, due to border-related trade barriers, and due to retail and wholesale distribution costs ($2.7 = 1.21*1.44*1.55$).

An extensive test of the basic implications of NEG theory, with a special focus on trade costs, has been carried out by Redding and Venables (2004) in the spirit of the NEG tradition (Fujita et al. 1999). The main empirical motivation of their work is understanding the extent of barriers to trade (mainly related to transport) which impede income, wages, and prices from equalizing across countries: in a neoclassical flat world without trade costs, arbitrage would exhaust all possible profit opportunities. Many of the reasons why observed prices systematically diverge from place to place relate to the physical transportation of goods from production places up to consumption markets and to trade policies. The authors employ: (1) a gravity-like relationship for bilateral trade flows between countries to measure the proximity to final goods and factor markets; (2) the demand for factors of the representative firm in each country, given its market access and supplier access; (3) a price index, linking the prices of final goods to access to factor markets. The quantitative results are striking: access to the coast and open-trade policies predict increases in per capita income of over 20 %, while halving a country's distance from trade partners yields an increase of around 25 %.

The econometric exercise is also very relevant from a validation viewpoint: the main tenets of the NEG theory look corroborated by the empirical evidence and the estimated magnitude of the structural parameters are in line with the reasonable values used in the literature. Moreover, Redding and Venables (2004) find that (in line with Hummels and Levinsohn 1995) the geography of access to markets and sources of supply is a powerful explanator of cross-country variation in per capita income.

Lafourcade and Thisse (2008) review the main results of the NEG approach about the effects of trade costs on the *distribution* of economic activities across space. One key ingredient of the NEG models is a focus on the accessibility to dispersed markets driven by all sorts of spatial frictions. In recent decades, some of these frictions, notably transport and communication costs, are steadily decreasing in importance, so it makes sense to ponder what kind of spatial pattern may emerge because of these changes in the structure of costs. In the light of NEG models, the response pattern of distribution of economic activities following a transport cost decrease may well be non-monotonic, with dispersion occurring both at very low and very high cost levels.

While social welfare comparisons (Charlot et al. 2006) between agglomeration and dispersion are essentially inconclusive, Fujita and Thisse (2002) emphasize that additional growth driven by agglomeration may lead to a Pareto-dominant configuration because the rate of technological innovation is higher when an economy moves from dispersion to agglomeration, thus boosting growth. When growth effects are sufficiently strong, concerns about regional disparities may be a lesser concern.

A relevant component of transport costs is determined by the efficiency of physical infrastructure: this is especially true for landlocked countries. To estimate the importance of infrastructure in determining the level of trade flows, Limao and Venables (2001) build an econometric gravity model, incorporating geographical characteristics (the shortest distance between countries, whether they share a common border, whether they are landlocked, and whether they are islands) and infrastructure measures.

The results show that geographical distance in itself cannot explain a significant part of the variation in transport costs; rather, the elasticity of cost factor CIF/FOB with respect to an infrastructure index (computed using the inverse of the index of road, the percentage of paved road, railway densities and telephone lines per capita) is found to be a significant variable. A country improving its transport infrastructure from the median to the top 25th percentile would face a remarkable equivalent 2,358 km decrease in the distance to all its trading partners. Trade flows are also very responsive to changes in the efficiency of infrastructure: if a country faces a deterioration of infrastructure from the median to the 75th percentile, this translates into a reduction of trade volumes by around 28 %. Moreover, landlocked countries are found to have systematically higher transport costs and lower trade volumes, while being or trading with an island reduces transport costs.

The policy suggestions from the econometric exercise are clear: investment in physical infrastructure can boost significantly trade flows of landlocked countries because of the costs from (1) border delays, (2) uncertainty and delays resulting in higher insurance fees, and (3) direct charges by the transit country. This is especially true of African countries, the trade of which is concentrated at the subregional level because of poor port infrastructure.

3.1 *Exogenous Trade Costs*

Several extensions of the basic models have been carried out to check whether the fundamental implications of the NEG model continue to hold in analytically richer frameworks. Basically, the main results of the simpler models remain valid, with some important qualifications.

The first paper to look into the black box of trade costs in a NEG model with two countries is Martin and Rogers (1995). The authors study the impact of international and domestic transport infrastructure on industry relocation and agglomeration using an iceberg cost function which allows for different transport costs of goods delivered internally and internationally. No distinction is made between trade costs due to institutional and transport factors. When new capital is added under the form of transport infrastructure facilitating international trade, the increasing return mechanism drives the relocation of industries toward the richest country. On the contrary, an improvement of domestic transport infrastructure creates an incentive for firms to relocate in that country. Accordingly, the model has a neat policy implication for Europe: if authorities intend to promote income convergence across states, they should concentrate investment in transport infrastructure more at the domestic level rather than at the international level.

Bosker et al. (2010) note that, though the applied NEG works usually employ multi-national or multi-regional data, the theoretical priors are mostly derived from the basic two-region model. Moreover, these works explicitly assume heterogeneity in transport costs ($T_{ij} \neq T$) and various types of trade frictions—tariff and non-tariff trade barriers. This improper *reductio* is probably due to the fact that a general n regions NEG model is analytically intractable and its equilibrium relations are not known: assuming that the predictions of the two-regions model applies without qualifications to higher dimensional models is then largely unfounded.

To circumvent the obstacle of mathematical intractability, Bosker et al. (2010) use computational methods to investigate the properties of a multi-regional NEG model (Puga 1999) in which transport costs assume the form

$$T_{ij} = T_{ji} = D_{ij}^{d}\left(1 + b_f B_{ij}\right)$$

where D_{ij}^{δ} is the great-circle distance between region i's and region j's capital city, B_{ij} is a dummy for the two regions being in the same country, δ is the distance decay parameter and b measures the strength of border frictions. Obviously, if $i = j$, then $T_{ij} = 1$. In this specification trade costs increase because of distance, national borders, and nontariff barriers. With this formulation, it is possible to break the correspondence between the level of agglomeration and the spatial distribution of economic activities which is peculiar of the two-region or evenly-spaced model.

Assuming alternately full labor mobility or immobility, Redding and Venables (2004) simulate (1) the effect of a decrease in the decay factor, setting border effects to zero, and (2) the effect of a decrease in border effects, setting transport costs to

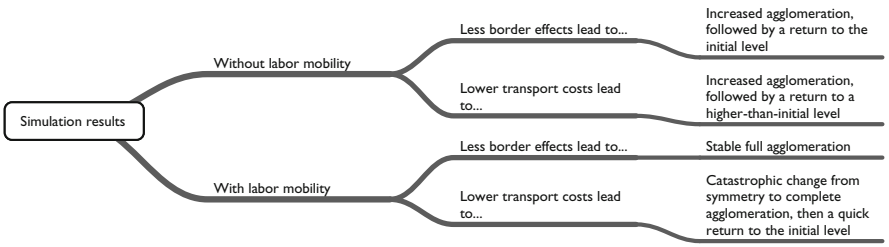


Fig. 2 Trade costs simulation in a multi-region model

zero: simulation results are reported in Fig. 2 and show that many features of the simple models carry over to a multi-region setting. In particular,

1. increased integration sparks agglomeration: a number of core regions attract activity from nearby regions while leaving some industrial activity in the peripheral regions;
2. agglomeration continues until complete specialization: agriculture in the periphery, industry in the core;
3. further integration reverses the agglomeration, gradually spreading activities from the core to nearby regions and then to peripheral regions.

In sum, in the case of an asymmetric spatial structure, increased integration drives catastrophic agglomeration, just like in the case of a symmetric structure; nonetheless, the same levels of agglomeration now can entail very different patterns of spatial distribution.

The most sophisticated theoretical contribution to date to the literature on trade costs in a multinational setting has been provided by Behrens et al. (2007): using a multi-country Dixit–Stiglitz trade model (Krugman 1980) enhanced with a rich graph-theoretic framework, the authors analyze how industry location and welfare respond to changes in transport and non-transport frictions. They are able to show that reductions in transport frictions occurring at *links around which the spatial network is locally a tree* are Pareto welfare improving.

The model can assess the impact of both transport and non-transport costs. On the one hand, changes in transport frictions, which are usually not origin–destination specific, may have predictable local effects. On the other hand, non-transport costs impact on the whole economy, given the changing incentives to the location of firms in the global economy: these costs are usually origin–destination specific and their changes are difficult to predict clearly. The graph structure of the model permits an appreciation of the fact that non-transport costs indirectly connect any pair of countries; whereas transportation occurs along specific routes so that unconnected and far-away countries are not impacted directly. The function of total trade frictions is

$$\tau_{ij} = \epsilon_{ij}(1 + \rho_{ij})$$

where ρ_{ij} is the ad valorem tariff equivalent of non-transport frictions between countries i and j, and ϵ is the transport friction between the same countries.

The authors use the results of the model to analyze European integration policies. A surprising implication they find is that market integration, like that taking place in Europe, may damage not only excluded countries but also small or far-off member countries with small markets. This drives the relocation of market-size dependent services toward the incumbent countries.

3.2 Endogenous Trade Costs

Another strand of literature explicitly questions the standard assumption of iceberg costs, i.e., transport costs entailing a fixed percentage of traded goods wasted along the way to its final destination.

Behrens et al. (2006) note that it is customary in the NEG literature to assume that transport costs are independent of trade volumes. Nonetheless, this is hardly the case in reality since economies of scale in transport are widespread. For example, infrastructures available in hubs are highly specialized and allow for low costs per unit on high volumes. Moreover, transport diseconomies are also relevant in the case of road transportation due to congestion. When shipping costs vary in response to traded volumes, trade costs become endogenous. Behrens et al. (2006) address this issue in a NEG model with two countries, each partitioned in two regions. Shipping costs per traded unit are functions of the volumes of international trade. However, there is no explicit modeling of the microeconomics of transport and just a reduced form relation for the trade cost is assumed of the type

$$T = f\left(X\left(\lambda_i, \lambda_j\right)\right) \tag{18}$$

where X is the equilibrium volume of international trade and λ_k is the fraction of skilled workforce located in the $k \in \{i, j\}$ country. Using a Taylor expansion, the function is approximated by the expression

$$T \approx \{1 - \xi[T_j(1 - \lambda_j) + T_i(1 - \lambda_i)]\} \tag{19}$$

where ξ measures the degree of density economies and T_k is the transport cost within the $k \in \{i, j\}$ country. The key finding is that agglomeration or dispersion of industry within a country may be driven by the geography of the other country through the channel of trade. This result is striking since it emphasizes that internal agglomeration may be responsive to the internal structure of the trading partner. Moreover, the authors find that density economies can lead to multiple equilibria and complete agglomeration in both countries, while diseconomies drive a smooth agglomeration process toward a unique stable equilibrium.

The apparently innocuous assumption of symmetric transport costs $T_{ij} = T_{ji}$ can be removed when the so-called *backhaul problem*, a very specific and empirically

relevant feature of the transport sector, is taken into account. Since trade flows are unevenly distributed across geographical space, ships and other carriers going from a net exporting country to a net importing country will leave with a full load and will return with insufficient cargo. Though the cost of physical transportation of a good from i to j is equal to the cost of delivering it from j to i, nonetheless the full opportunity costs diverge. The strong demand for shipping from i to j by exporting firms located in i is contrasted by a weak demand for shipping goods the other way around; accordingly, the full freight rates will be different, with the cost of freight from i to j being higher than the cost of freight from j to i. For example, air freights from China to North America cost \$3/3.50 per kilogram, whereas freights from North America to China cost \$0.3/0.4 per kilogram. While the backhaul problem has already attracted the attention of applied economists and operation research scientists (see, for example, Anderson and Wilson 2008; Dejax and Crainic 1987), its full implications for economic geography models have not been fully grasped so far. The implications of the asymmetries in freight rates are far-reaching since they are able to contrast the tendencies to agglomeration and specialization of industry.

Behrens and Picard (2011) explore the role of transport costs in a NEG model with two regions (Ottaviano et al. 2002), assuming a perfectly competitive transport sector and price setting which accounts for backhauling. In their model, two alternative specifications are employed: the *footloose capital* model and the *core–periphery* model. In both specifications they invariably find that a more even spatial distribution of firms and production prevails when freight rates are determined by the endogenous supply-and-demand mechanism than when they are taken as exogenous.

The presence of endogenous freight rates weakens the role of the Home Market Effect (HME). In the footlose capital model, endogenous freight rates lead to activity dispersion up to the point of no HME. In the core–periphery model they lead to multiple stable equilibria in which full agglomeration and full dispersion may be stable equilibria. Relaxing the assumptions of perfectly competitive markets in transport or assuming that firms incur in an additional loading/handling fixed cost for each unit of shipped good does not alter the essential qualitative results of the model.

4 Some Reflections on Policy Issues

NEG models have often been used to assess the impact of various policy measures including tariffs, free-trade agreements, customs unions, taxes, subsidies, public expenditures on infrastructure, transport systems and research and development on the regional distribution of economic activity and welfare (see for an overview Baldwin et al. 2003). However, only few papers have explicitly addressed the question whether the properties of the decentralized market equilibria are socially desirable and how an optimal policy should be designed. This is astonishing since the very core of a prototype NEG model structure involves several inefficiencies—in

addition to the monopolistic distortion, the change in the locally available amounts of productive factors involves pecuniary externalities that are welfare-relevant in the given context of imperfect competition. A central stream of papers in this field—Ottaviano and Thisse (2001, 2002), Ottaviano et al. (2002), Tabuchi and Thisse (2002) and more recently Pflüger and Südekum (2008) (linking their analysis to Helpman 1998)—introduce a specific variant of a social planner, in particular one which imposes marginal cost pricing, uses lump-sum transfers to pay for losses involved and chooses the spatial allocation of the mobile factor such that the sum of the indirect utilities is maximized.[6] These authors derive parameter ranges (in particular for trade freeness) for which the symmetric and the core–periphery (CP) equilibrium are welfare maximizing and they show that those ranges do not necessarily coincide with the parameter ranges for which the respective type of equilibrium is the stable outcome of the decentralized market processes.[7] Ottaviano and Thisse (2001), Ottaviano et al. (2002), Pflüger and Südekum (2008) interpret this divergence as opening up room for regional policy interventions without specifying them in detail; Ottaviano et al. (2002) are a bit more explicit and argue for restricting factor mobility when market processes would produce over-agglomeration, i.e., agglomeration in a parameter range within which the symmetric equilibrium exhibits a higher social welfare. Alternatively, they suggest interregional transfers to compensate the periphery in a similar vein to Tabuchi and Thisse (2002, p. 173) who also mention interregional income transfers. However, none of these studies explicitly derives policy recommendations on the basis of the model analysis.

A recent contribution by Grafeneder-Weissteiner et al. (2012), on the other hand, explicitly introduces a social planner into an NEG model which directly chooses quantities (allocation of productive factors, allocation of outputs, given preferences and given technology) and does not impose any price mechanism. They derive an optimal subsidy/taxation scheme in the sense that its implementation would adjust the solution reached in a decentralized market economy in the symmetric equilibrium, to the solution of the social planner problem. Not surprisingly, it turns out that the optimal policy is a sales subsidy financed by a lump-sum tax that results in marginal cost pricing. In addition, they show that implementing this optimal policy into the decentralized market economy may actually destroy stability of the symmetric equilibrium in the decentralized economy. Thus starting from a symmetric equilibrium (corresponding to equity considerations), the attempt to increase economic efficiency actually triggers a dynamic process that leads away from this equitable situation.

However, it is not only the—so far mostly neglected—dynamic implications of policy intervention that make it difficult to conduct an adequate policy and welfare

[6] In addition, Ottaviano et al. (2002) and Pflüger and Südekum (2008) also analyse a second-best solution for the social planner, i.e., a solution in which the social planner is assumed not to change market prices, but only to optimally choose the factor location.

[7] Note, however, that they do not consider stability issues in the social planner solution.

analysis in these frameworks. Also, the lack of a missing representative agent in many NEG models and thus distributive issues further complicate the analysis. In particular, the utility level of the immobile workers left behind in the periphery is lower than the utility level in the core region. The use of an appropriate social welfare function thus becomes a fundamental issue. Ottaviano et al. (2002) as well as Tabuchi and Thisse (2002) explicitly analyse the welfare position of different groups. In such a situation, the utilization of a social welfare function is not unproblematic. Charlot et al. (2006), pointing out that the simple utilitarian social welfare function actually reflects indifference to inequality, suggest using the more general CES specification that is able to represent a wide range of societal attitudes toward inequality. In addition, they apply compensation criteria (cf. Robert-Nicoud 2006; Kranich 2009) in order to directly rank the two possible market outcomes, namely the symmetric, dispersed equilibrium and a CP equilibrium. They show that the result heavily depends on attitudes toward inequality. For plausible parameter values they show that the market might lead to over-agglomeration. Again, policy implications are not at their focus. Similar to the papers reviewed above, they cautiously recommend not to intervene in agglomerative processes, but to use interregional transfers to compensate ex post for the lower utility levels in the periphery. The reason given for that position is worth quoting: "we find it hard to recommend a move from a stable equilibrium, such as agglomeration, to a socially preferred unstable equilibrium, such as dispersion." (Charlot et al. 2006, p. 343).

5 Conclusions

Since not long ago, NEG models were typically confined to two regions. The analytics of models involving only two regions is already surprisingly complex; this might be one of the reasons, why not much effort was dedicated in extending the number of regions. However, as noted in the introduction, a two regions framework might conceal indirect effects that can only occur in a model with more than two countries.

Thus, in order to come closer to an applicability of NEG models to assess real world policy designs, this extension is of utmost importance. The present review intends to cover the already existing literature, focusing on the footloose entrepreneur model. What are the crucial points?

First, with more than two regions, the structure of the transport cost network connecting the countries/regions involved becomes an issue—are transport cost the same between all countries? or do asymmetries exist, such as a hub and spoke structure, in which some regions assume a central position in the sense of providing an easy access to the other regions. Therefore, we dedicated a separate section to the question how to model and measure transport cost.

Second, in reviewing the model structures it turned out that the specification of preferences and thus demand functions is decisive for the model outcomes. Typically, NEG models involve iso-elastic demand functions—which in turn implies

a constant mark-up for the pricing decision. The competition effect that is the central dispersion force works only through the market share (the size of the market niche). We compared this standard modelling framework to one that uses a linear demand function (as introduced by Ottaviano et al. 2002). With a linear demand function, the competition effect is reinforced—more competitors not only reduce the market share but reduces the mark up as well. Not surprisingly, with a linear demand function, the competition effect can prevail and reducing the external transport cost might lead to more dispersion within a country (contrary to the results obtained with an iso-elastic demand function).

Third, when applying NEG models to policy issues, particular attention has to be paid to distributive issues (since the assumption of a representative agent is not always applicable) and to the dynamic framework. In a separate section, we covered the emerging strand of literature.

References

Ago, T., Isono, I., & Tabuchi, T. (2006). Locational disadvantage of the hub. *The Annals of Regional Science, 40*(4), 819–848.

Akamatsu, T., & Takayama, Y. (2009). *A simplified approach to analyzing multi-regional core–periphery models*. Technical report, University Library of Munich, Munich, Germany.

Akamatsu, T., Takayama, Y., & Ikeda, K. (2012). Spatial discounting, fourier, and racetrack economy: A recipe for the analysis of spatial agglomeration models. *Journal of Economic Dynamics and Control, 36*(11), 1729–1759.

Alonso-Villar, O. (2001). Large metropolises in the third world: An explanation. *Urban Studies, 38*(8), 1359–1371.

Anderson, J. E., & van Wincoop, E. (2004). Trade costs. *Journal of Economic Literature, 42*(3), 691–751.

Anderson, S. P., & Wilson, W. W. (2008). Spatial competition, pricing, and market power in transportation: A dominant firm model. *Journal of Regional Science, 48*(2), 367–397.

Baldwin, R., Forslid, R., Martin, P., Ottaviano, G., & Robert-Nicoud, F. (2003). *Economic geography and public policy*. Princeton, NJ: Princeton University Press.

Baldwin, R. E. (1999). Agglomeration and endogenous capital. *European Economic Review, 43*(2), 253–280.

Behrens, K. (2011). International integration and regional inequalities: How important is national infrastructure? *The Manchester School, 79*(5), 952–971.

Behrens, K., Gaigné, C., Ottaviano, G. I., & Thisse, J.-F. (2006). How density economies in international transportation link the internal geography of trading partners. *Journal of Urban Economics, 60*(2), 248–263.

Behrens, K., Lamorgese, A. R., Ottaviano, G. I., & Tabuchi, T. (2007). Changes in transport and non-transport costs: Local vs global impacts in a spatial network. *Regional Science and Urban Economics, 37*(6), 625–648.

Behrens, K., & Picard, P. M. (2011). Transportation, freight rates, and economic geography. *Journal of International Economics, 85*(2), 280–291.

Bosker, M., Brakman, S., Garretsen, H., & Schramm, M. (2010). Adding geography to the new economic geography: Bridging the gap between theory and empirics. *Journal of Economic Geography, 10*(6), 793–823.

Brülhart, M., Crozet, M., & Koenig, P. (2004). Enlargement and the EU periphery: The impact of changing market potential. *The World Economy, 27*(6), 853–875.

Castro, S. B., Correia-da Silva, J., & Mossay, P. (2012). The core–periphery model with three regions and more. *Papers in Regional Science, 91*(2), 401–418.

Charlot, S., Gaigne, C., Robert-Nicoud, F., & Thisse, J.-F. (2006). Agglomeration and welfare: The core–periphery model in the light of Bentham, Kaldor, and Rawls. *Journal of Public Economics, 90*(1), 325–347.

Commendatore, P., & Kubin, I. (2013). A three-region new economic geography model in discrete time: Preliminary results on global dynamics. In *Global analysis of dynamic models in economics and finance* (pp. 159–184). Berlin Heidelberg: Springer.

Commendatore, P., Kubin, I., Petraglia, C., & Sushko, I. (2012). *Economic integration and agglomeration in a customs union in the presence of an outside region*. WU Economics Working Paper 146, WU Vienna University of Economics and Business, Vienna.

Commendatore, P., Kubin, I., Petraglia, C., & Sushko, I. (2014). *Regional integration, international liberalisation and the dynamics of industrial agglomeration*. WU Economics Working Paper 164, WU Vienna University of Economics and Business.

Crozet, M., & Koenig Soubeyran, P. (2004). EU enlargement and the internal geography of countries. *Journal of Comparative Economics, 32*(2), 265–279.

Dejax, P. J., & Crainic, T. G. (1987). A review of empty flows and fleet management models in freight transportation. *Transportation Science, 21*(4), 227–248.

Forslid, R. (2004). *Regional policy, integration and the location of industry in a multiregion framework*. CEPR Discussion Paper No. 4630, London: Centre for Economic Policy Research.

Forslid, R., & Ottaviano, G. (2003). An analytically solvable core–periphery model. *Journal of Economic Geography, 3*(3), 229–240.

Fujita, M., Krugman, P. R., & Venables, A. J. (1999). *The spatial economy: Cities, regions and international trade* (Vol. 213). Cambridge Massachusetts: The MIT Press.

Fujita, M., & Thisse, J.-F. (2002). *Economics of agglomeration*. Cambridge, MA: Cambridge University Press.

Fujita, M., & Thisse, J.-F. (2009). New economic geography: An appraisal on the occasion of Paul Krugman's 2008 Nobel Prize in economic sciences. *Regional Science and Urban Economics, 39*(2), 109–119.

Gaspar, J., de Castro, S. B. S. D., & Silva, J. C. D. (2013). *The footloose entrepreneur model with 3 regions*. FEP Working Paper 496, Universidade do Porto, Faculdade de Economia do Porto.

Grafeneder-Weissteiner, T., Kubin, I., Prettner, K., Fürnkranz-Prskawetz, A., & Wrzaczek, S. (2012). *Coping with inefficiencies in a new economic geography model*. Working Paper, Vienna Institute of Demography (VID) of the Austrian Academy of Sciences in Vienna, Vienna.

Helpman, E. (1998). The size of regions. In D. Pines, E. Sadka, & I. Zilcha (Eds.), *Topics in public economics: Theoretical and applied analysis*. Cambridge, MA: Cambridge University Press.

Hummels, D., & Levinsohn, J. (1995). Monopolistic competition and international trade: Reconsidering the evidence. *The Quarterly Journal of Economics, 110*(3), 799–836.

Ikeda, K., Akamatsu, T., Kono, T. (2012). Spatial period-doubling agglomeration of a core–periphery model with a system of cities. *Journal of Economic Dynamics and Control, 36*(5), 754–778.

Ikeda, K., Murota, K., Akamatsu, T., Kono, T., Takayama, Y., Sobhaninejad, G., & Shibasaki, A. (2010). *Self-organizing hexagons in economic agglomeration: Core–periphery models and central place theory*. Technical report 28, Department of Mathematical Informatics, University of Tokyo, Tokyo.

Kranich, J. (2009). Agglomeration, innovation and international research mobility. *Economic Modelling, 26*(5), 817–830.

Krugman, P. (1980). Scale economies, product differentiation, and the pattern of trade. *The American Economic Review, 70*(5), 950–959.

Krugman, P. (1991). Increasing returns and economic geography. *The Journal of Political Economy, 99*(3), 483–499.

Krugman, P. (1993). The hub effect: Or, threeness in international trade. In M. Fujita, P. Krugman, & A. J. Venables (Eds.), *The spatial economy: Cities, regions, and international trade*. Cambridge, MA: MIT Press.

Krugman, P., & Elizondo, R. L. (1996). Trade policy and the third world metropolis. *Journal of Development Economics, 49*(1), 137–150.

Krugman, P., & Venables, A. J. (1995). Globalization and the inequality of nations. *The Quarterly Journal of Economics, 110*(4), 857–880.

Lafourcade, M., & Thisse, J.-F. (2008). *New economic geography: A guide to transport analysis*. PSE Working Paper 2008-2, Paris-Jourdan Sciences Economiques Laboratoire D'economie Appliquee – INRA.

Limao, N., & Venables, A. J. (2001). Infrastructure, geographical disadvantage, transport costs, and trade. *The World Bank Economic Review, 15*(3), 451–479.

Martin, P., & Rogers, C. A. (1995). Industrial location and public infrastructure. *Journal of International Economics, 39*(3), 335–351.

Monfort, P., & Nicolini, R. (2000). Regional convergence and international integration. *Journal of Urban Economics, 48*(2), 286–306.

Monfort, P., & Van Ypersele, T. (2003). *Integration, regional agglomeration and international trade*. CEPR discussion paper 3752 Economic Policy, London: Centre for Economic Policy Research. .

Østbye, S. (2010). Regional policy analysis in a simple general equilibrium model with vertical linkages. *Journal of Regional Science, 50*(3), 756–775.

Ottaviano, G., Tabuchi, T., & Thisse, J.-F. (2002). Agglomeration and trade revisited. *International Economic Review, 43*(2), 409–435.

Ottaviano, G. I., & Thisse, J.-F. (2001). On economic geography in economic theory: Increasing returns and pecuniary externalities. *Journal of Economic Geography, 1*(2), 153–179.

Ottaviano, G. I., & Thisse, J.-F. (2002). Integration, agglomeration and the political economics of factor mobility. *Journal of Public Economics, 83*(3), 429–456.

Paluzie, E. (2001). Trade policy and regional inequalities. *Papers in Regional Science, 80*(1), 67–85.

Pflüger, M. (2004). A simple, analytically solvable, chamberlinian agglomeration model. *Regional Science and Urban Economics, 34*(5), 565–573.

Pflüger, M., & Südekum, J. (2008). A synthesis of footloose-entrepreneur new economic geography models: When is agglomeration smooth and easily reversible? *Journal of Economic Geography, 8*(1), 39–54.

Puga, D. (1999). The rise and fall of regional inequalities. *European Economic Review, 43*(2), 303–334.

Puga, D., & Venables, A. J. (1996). The spread of industry: Spatial agglomeration in economic development. *Journal of the Japanese and International Economies, 10*(4), 440–464.

Puga, D., & Venables, A. J. (1997). Preferential trading arrangements and industrial location. *Journal of International Economics, 43*(3), 347–368.

Puga, D., & Venables, A. J. (1999). Agglomeration and economic development: Import substitution vs. trade liberalisation. *The Economic Journal, 109*(455), 292–311.

Redding, S., & Venables, A. J. (2004). Economic geography and international inequality. *Journal of International Economics, 62*(1), 53–82.

Robert-Nicoud, F. (2006). Agglomeration and trade with input–output linkages and capital mobility. *Spatial Economic Analysis, 1*(1), 101–126.

Roos, M. W. (2005). How important is geography for agglomeration? *Journal of Economic Geography, 5*(5), 605–620.

Tabuchi, T., & Thisse, J.-F. (2002). Taste heterogeneity, labor mobility and economic geography. *Journal of Development Economics, 69*(1), 155–177.

Venables, A. J. (1996). Equilibrium locations of vertically linked industries. *International Economic Review, 37*(2), 341–359.

Venables, A. J. (2006). *Shifts in economic geography and their causes*. Discussion paper 767, Centre for Economic Performance, London School of Economics and Political Science, London, UK.

Villar, O. A. (1999). Spatial distribution of production and international trade: A note. *Regional Science and Urban Economics, 29*(3), 371–380.

Wang, J., & Zheng, X.-P. (2013a). Industrial agglomeration and dispersion in gate and hinterland regions. *The Ritsumeikan Economic Review, 62*(1), 39–60.

Wang, J., & Zheng, X.-P. (2013b). Industrial agglomeration: Asymmetry of regions and trade costs. *Review of Urban & Regional Development Studies, 25*(2), 61–78.

Wiggins, S. (2003). *Introduction to applied nonlinear dynamical systems and chaos* (2nd ed.). New York, NY: Springer.

Parametric Models in Spatial Econometrics: A Survey

Diana A. Mendes and Vivaldo M. Mendes

Abstract The main purpose of this chapter is to review the parametric spatial econometric models that can be applied to regional economics. Spatial econometric methods are based on regression analysis applied to cases where spatial interactions and spatial structures are fundamental characteristics of the process under discussion. The review presented here outlines the basic terminology, the spatial data dependence, the specification of spatial effects, and some basic spatial regression models, i.e., the spatial autoregressive (SAR) model (or spatial lag model), the spatial error model (SEM), the spatial Durbin model (SDM) and the general spatial models—SAC and SARMA. The maximum likelihood estimation for SAR and SEM models it is also presented with some detail.

In the context of the European Union, we should emphasize several empirical works in the particular areas of urban economics, economic growth and productivity, and studies dealing with agglomeration and externalities (spillovers). We provide a brief survey of some of the results obtained in these particular areas.

1 Introduction

The main purpose of this paper is to provide a review of the basic tools and models from spatial data analysis and spatial econometrics. The reader can find several books on spatial econometrics, which by their nature and length can provide a somewhat more detailed view and scope of the particular subject of this survey, e.g., Anselin (1988), Getis et al. (2004), Arbia (2006), LeSage and Pace (2009), among others. Our main purpose, however, is to provide a brief survey of the main spatial parametric models and some applications of them to EU spatial data.

Applied work in regional science frequently deals with sample data that is collected with reference to location and measured as points in space. Generally, there are two problems that arise when sample data has a locational component:

D.A. Mendes (✉) • V.M. Mendes
Department of Quantitative Methods for Business and Economics, ISCTE-IUL and BRU-IUL, Avenida das Forças Armadas, 1649-026 Lisbon, Portugal
e-mail: diana.mendes@iscte.pt; vivaldo.mendes@iscte.pt

P. Commendatore et al. (eds.), *Complexity and Geographical Economics*, Dynamic Modeling and Econometrics in Economics and Finance 19,
DOI 10.1007/978-3-319-12805-4_3

(a) spatial dependence exists among the observations, and (b) spatial heterogeneity occurs in the relationships we are modeling. The standard econometric tools that are available are not suitable to deal with these two problems in an efficient manner, as they violate the traditional Gauss–Markov assumptions used in regression modeling. With regard to the spatial dependence among the observations, we should recall that the Gauss–Markov assumption takes that the explanatory variables are fixed in repeated sampling. Spatial dependence clearly violates this assumption and this leads to the need for alternative estimation approaches. Similarly, spatial heterogeneity violates the Gauss–Markov assumption that a single linear relationship exists across the sample data observations. If the relationship varies, as we move across the spatial data sample, alternative estimation procedures are needed to successfully model this type of variation and to draw appropriate inferences. When the classical methods fail, the use of spatial econometrics (SE) is recommended, since SE allows us to explain the agglomeration processes and uneven spatial distribution of economic activities across regions.

Historically, in the early 1970s, spatial econometrics emerged as a proper and separate subject within the field of econometrics in Europe, to deal with sub-country data in regional econometric models (Hordijk and Paelinck 1976; Paelinck and Klaassen 1979). During this first stage of development of spatial econometrics, interest was focused on testing for residual spatial autocorrelation (using Moran's I), the specification of spatial models, the basic estimation methods, model discrimination and specification testing, as well as some initial work on space-time models. Our review will cover these topics as they are nowadays considered basic topics in spatial econometrics, and also because we can find several applications for them in most empirical works dealing with spatial data.

More recently, spatial econometric methods have increasingly become more frequently applied in the traditional fields of economics, such as in international economics, agglomeration economics and New Economic Geography, labor economics, public economics, and agricultural and environmental economics (Gorter et al. 2005; Fingleton 2000, 2004; LeSage and Fischer 2008; Basile 2009; Rey and LeGallo 2009). These new applications have raised attention to new problems and challenges which can be grouped in four main domains: spatial bias, spatial specification, spatial estimation, and spatial complexity. These new problems have led to new scientific research areas in spatial econometrics which have become currently known as the "non-standard spatial econometrics" (see for instance Griffith and Paelinck 2011).

This short survey is organized as follows. Section 2 deals with spatial neighbors and weights, while Sect. 3 discusses with some detail spatial dependence. Section 4 presents the basic spatial regressive/autoregressive model (SAR, SEM, SAC, SARMA), spatial effects, maximum likelihood (ML) estimation method and some applications for empirical spatial data in the context of European Union. The final section makes a quick review of the currently available software packages for spatial econometrics.

2 Spatial Neighbors and Spatial Weights

Spatial data samples represent observations that are associated with points or regions, have coordinate values and a system of reference for these coordinates. Usually we can find four different types of data samples (models) that can be distinguished as following: point sample (meaning a single point location), line sample (that is a set of ordered points connected by straight line segments), polygon or areal sample (that is, geometric figures delimited by one or more boundary lines, possibly containing holes), and grid sample which is a collection of points or rectangular cells, organized in a regular lattice. The first three are vector data models and are exactly defined, while the fourth one is a raster data model, representing continuous surfaces by using a regular tessellation.[1] All spatial data are characterized by positional information or spatial data, and some of them can be extended also by some attributes, e.g., for example spatiotemporal data.

In this note we are more concerned with polygonal or areal data in the context of several standard spatial econometric models. Our setting is a data-set of spatial regions and we shall assume that the data set contains a single observation on each region (i.e., an observation at a single point in time), comprising a spatial cross-section.

The essence of spatial analysis is that "space matters", such that what happens in one region is related to what happens in other neighboring regions. Spatial autocorrelation is the correlation among values of a single variable strictly attributable to their relatively close locational positions on a two-dimensional surface, causing a violation of the independent observations assumption of classical statistics. Spatial autocorrelation exists because real-world phenomena are typified by orderliness, pattern, and systematic concentration, rather than randomness. This has been made more precise in what Tobler (1970) refers to as the First Law of Geography: "Everything is related to everything else, but near things are more related than distant things."

Before analyzing spatial dependence and heterogeneity, we should beware of the quantification of the locational aspects of the sample data. The location of data in the Cartesian plane, given by latitude and longitude, is a very important source of information which allow us to calculate distances from any point in space, or the distance between different locations. Another source of locational information is given by contiguity, reflecting the relative position in space of one regional unit of observation with respect to other such units. Measures of contiguity are based on the knowledge of the size and shape of the observational units illustrated on a map. From this graphical representation, we can observe which units are neighbors (have common borders) or situated in a reasonable proximity to each other. Usually, neighboring units exhibit a higher degree of spatial dependence than units placed far away from each other.

[1] Regular tessellation is a pattern made by repeating a regular polygon.

The weights matrix can be considered the central part of spatial econometrics models since it defines the strength of the interaction among spatial areal units. A weights matrix is used to represent which of these spatial units (regions, countries) are neighbors to each other. There are several ways to specify a spatial weights matrix, such as distance, contiguity, economic distance, among others (see LeSage and Pace 2009, for example). The choice of the weights matrix is a delicate step in spatial analysis and can influence the significance of testing procedures. The guiding principle in selecting a definition should be the nature of the problem being modeled, and perhaps particular additional non-sample information which may be available.

The first step in creating spatial weights is to define which relationships between observations are to be given a nonzero weight, that is, to choose the neighbor criterion to be used, while the second step is to assign weights to the identified neighbor links. The basis for most models is an indicator of whether one region is a spatial neighbor of another, or equivalently, which regions are neighbors of a given region. The most common type is the binary weights matrix. In this case, for a given sample of size n, we define a square symmetric matrix $(n \times n)$ where any entry w_{ij}, $i = 1, \ldots, n,\ j = 1, \ldots, n$, is given by

$$w_{ij} = \begin{cases} 1 & \text{if regions } i \text{ and } j \text{ are neighbors (spatially related)} \\ 0 & \text{otherwise} \end{cases}$$

By convention, the diagonal elements of this weights matrix are set to zero, since spatial econometric models assume that each spatial unit does not consider itself to be its own neighbor. Violating this assumption can result in a considerably more complex model that cannot be easily interpreted.

There exists a large number of ways to construct (first order) contiguity weights matrices (see for instance LeSage 1998; Bivand et al. 2008). These include:

- Rook contiguity: two regions are neighbors if they share a common border (on any side). In practice, we consider a distance threshold and say that two regions share a common border if that border is longer than the distance threshold. Define $w_{ij} = 1$ for regions that share a common side with the region of interest.
- Bishop continuity: two regions are spatial neighbors if they meet at some point. This is the spatial analog of two elements of a graph meeting at a vertex. In practice, we consider a distance threshold and say that two regions are neighbors if their common border is shorter than the distance threshold. Define $w_{ij} = 1$ for regions that share a common vertex with the region of interest.
- Queen contiguity: this is the union of Rook and Bishop contiguity. Two regions are neighbors in this sense if they share any part of a common border, no matter how short, and we consider in this case that $w_{ij} = 1$.

We can also define second order measures of contiguity: that is, we count as neighbors the regions sharing a border with a first-order neighbor according to each of the criteria listed above.

Another approach to define weights matrices is distance-based, where we determine neighbors based on some distance threshold d. For areal data (regions), the geometric centroid of each polygon is used to calculate this distance. For instance, let d_{ij} be the distance between (centroids of) regions i and j. We define

$$w_{ij} = \begin{cases} 1 \text{ if } d_{ij} < d \\ \\ 0 \text{ otherwise} \end{cases}$$

for a pre-specified d. In other words, regions i and j are spatial neighbors when they are within d distance units of one another. This approach can be expanded in several ways, for example, one could use different distances, or different weights for computing the centroids.

The K-nearest-neighbors approach also uses the distance threshold concept but chooses a threshold such that each observation has exactly "K" neighbors. With a weights matrix based on the 2-nearest neighbors criterion, the two closest regions (countries), based on centroid-to-centroid distance, would be considered neighbors of the regions (countries) of interest.

It is quite frequent to use, in practice, some transformed weights matrices. The most common transformation is called "row-standardization", where the sum of the rows of the neighbors matrix are assumed to be equal to unity. Let $\tilde{W}$ with elements $\tilde{w}_{ij}$ be a spatial neighbors matrix (binary type). In order to row-standardize this matrix, we divide each element in a row by the sum of the elements in the row. Thus, we obtain, a new spatial weights matrix W; with element w_{ij} defined by

$$w_{ij} = \frac{\tilde{w}_{ij}}{\sum \tilde{w}_{ij}}$$

or, more precisely, since region i may have no neighbors (in this case, an island), we will have

$$W_{ij} = \tilde{w}_{ij} \,/\, \max\left(1, \sum \tilde{w}_{ij}\right)$$

In practice, most regional-type research begins with the simple $0/1$ row-standardized contiguity matrix. If the given data represents several regions, then, it will be quite difficult to work with large size arrays. In this case, it is recommended to only keep track of the elements that are not zero, as in this way we save considerable space. That means that we further proceed with a sparse-matrix representations. Instead of working with the entire matrix, we store all non-zero elements of the matrix into a linear array and provide auxiliary arrays to describe the locations of the non-zero elements in the original matrix.

3 Spatial Dependence (Autocorrelation)

Spatial dependence in a set of sample data observations refers to the fact that one observation associated with the location i depends on other observations at locations j, with $i \neq j$. Formally, we have that: $y_i = f\left(y_j\right),\ i = 1, \ldots, n, \quad j \neq i$.

Notice that we allow the dependence to be among several observations, as the index i can take on any value from $i = 1, \ldots, n$. Why would we expect sample data observed at one point in space to be dependent on values observed at other locations? There are two reasons commonly presented to justify thus. First, data collection of observations associated with spatial units might reflect measurement error. This would occur if the administrative boundaries for collecting information do not accurately reflect the nature of the underlying process generating the sample data. A second reason we would expect spatial dependence is that the spatial dimension of economic activity may truly be an important aspect of a modeling problem. Regional science is based on the premise that location and distance are important forces at work in human geography and market activity. All of these notions have been formalized in the theory of regional science that relies on notions of spatial interaction and diffusion effects, hierarchies of place and spatial spillovers.

Autocorrelation literally means that a variable is correlated with itself. Spatial autocorrelation is something like temporal autocorrelation, but somehow more sophisticated. The simplest definition of autocorrelation states that those pairs of areal units which are close to each other are more likely to have values that are more similar, and pairs of areal units far apart from each other are more likely to have values that are less similar. The spatial structure of the data refers to any pattern that may exist. Gradients or clusters are examples of spatial structures that are positively correlated, whereas negative correlation exhibits patterns where subjects appear to repulse each other. The absence of autocorrelation (random patterns) implies that data observations are statistically independent.

Formally, if y_i and y_j are realizations of a random variable y indexed by spatial locations, then we have spatial autocorrelation if

$$Cor(y_i, y_j) = E(y_i y_j) - E(y_i)E(y_j) \neq 0$$

The most common test for the existence of spatial autocorrelation is due to Patrick Moran, and is usually referred as Moran's-I Test. This statistic is defined for a particular data (or residual) vector y by Bivand et al. (2008):

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} \left(y_i - \overline{y}\right)\left(y_j - \overline{y}\right)}{\sum_i \left(y_i - \overline{y}\right)^2}$$

where n is the number of cases, y_i is the variable value at a particular location, y_j is the variable value at another location, $\overline{y}$ is the mean of y, and w_{ij} represents elements of the spatial weights matrix and measure the spatial proximity between region i and j. Similar to the correlation coefficient, it varies between -1 and $+1$,

so a high I value indicates positive autocorrelation and a low I value indicates negative autocorrelation. A p-value of this test statistic of 0.05 or lower indicates that spatial autocorrelation is present in some form. The computation of the Moran's test is relative to a given choice of the spatial weights W. If, in fact, the pattern of spatial autocorrelation is generated by a different set of weights, then the test can give spurious results. This test is already available in a number of different software packages including ArcGIS, R, Matlab.

Another test for spatial autocorrelation frequently found in the literature is the Geary's C statistic, defined as

$$C = \frac{n-1}{2\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (y_i - \overline{y})^2}{\sum_i (y_i - \overline{y})^2}.$$

The values of this statistic typically range between 0 and 2. An uncorrelated process has an expected value of $C = 1$, and values less than 1 indicate positive spatial autocorrelation, while values greater than 1 indicate negative autocorrelation.

Notice that this is not the inverse of the Moran's statistic as C approaches 0 and I approaches 1 when similar values are clustered, C approaches 2 and I approaches -1 when dissimilar values tend to cluster, and high values of C measures correspond to low values of I. Moran's statistic is usually said to be a measure of global spatial autocorrelation, while Geary's statistic is more sensitive to local autocorrelation. Interaction is not the cross-product of the deviations from the mean, but from the deviations in intensities of each observation location with one another.

The local Moran's I_i test,

$$I_i = n (y_i - \overline{y}) \frac{\sum_j w_{ij} \left(y_j - \overline{y}\right)}{\sum_j (y_i - \overline{y})^2}$$

is usually used to detect if local autocorrelation exists around each region i and it is usually applied after the global Moran's I. This local test allow us to study the homogeneity of the regions and of the total area, and also to detect if there are local outliers that contribute to a significant global statistic.

Another underlying condition leading to the necessity of using spatial econometrics is spatial heterogeneity. This term refers to variation (that is, lack of spatial stability) in trait, event, or relationships over space (Anselin 2006). In the most general case we might expect a different relationship to hold for every point in space. It is frequently introduced simultaneously with spatial dependence and, in practice, the two terms can be difficult to be distinguish from each other. Spatial heterogeneity is also sometimes referred to as "sub-regional variation," or "first-order spatial effects."

4 Spatial Models

Spatial econometrics is the collection of econometric tools dealing with problems of spatial dependence and spatial heterogeneity (heteroskedasticity). In this section we lay out the basic spatial regressive/autoregressive models introduced to model cross-sectional spatial data samples by following Anselin (1988, 2006), and LeSage and Pace (2009).

Our starting point is the conventional linear-in-parameters cross-sectional model given by

$$y_i = X_i\beta + u_i,\ i = 1, \ldots, n$$

where y is the vector of n dependent variables, X is the vector of n independent (explanatory) variables, β is the vector of the associated parameters and $u \sim N\left(0, \sigma^2\right)$ is the classical error term. Notice that the dependent and independent variables are linearly related. This type of data generating process is typically assumed for linear regression models. Each observation has an underlying mean of $X_i\beta$ and a random component u_i. An implication of this, for situations where the observations i represent regions or points in space, is that observed values at one location (or region) are independent of observations made at other locations (or regions). The assumption of independence greatly simplifies models, but in spatial contexts this simplification seems strained. In contrast, spatial dependence reflects a situation where values observed at one location or region, say observation i, depend on the values of neighboring observations at nearby locations. This situation suggests a simultaneous data generating process, where the value taken by y_i depends on that of y_j and vice versa.

There are three main ways to introduce spatial information in a linear regression or autoregression model: (a) spatially lagged independent variable, (b) spatially lagged dependent variable and (c) spatially lagged error term. In what follows we review each one of these cases. Linearity is for ease of notation, and any nonlinear extensions available in econometric models are also available in the spatial framework.

4.1 SAR: The Spatial Autoregressive Model (Spatial Lag Model)

The first order autoregressive model (FAR) is given by

$$y = \lambda Wy + u$$
$$u \sim N\left(0, \sigma^2 I_n\right)$$

where y contains an $(n \times 1)$ vector of cross-sectional dependent variables, W is a known $(n \times n)$ spatial weights matrix, usually containing first-order contiguity relations or functions of distance, λ represents a regression parameter to be estimated and u denotes the stochastic disturbance in the relationship above. The parameter λ would reflect the spatial dependence inherent in the sample data, measuring the average influence of neighboring or contiguous observations on observations in the vector y. If we posit spatial dependence between the individual observations in the data sample y, some part of the total variation in y across the spatial sample would be explained by each observation's dependence on its neighbors. The term Wy is called a spatial lag, since it represents a linear combination of values of the variable y constructed from observations/regions which are neighbors to observation i. We use $N(0, \sigma^2 I_n)$ to denote a zero mean error process that exhibits constant variance σ^2, and zero covariance between observations. This results in the diagonal variance-covariance matrix $\sigma^2 I_n$, where I_n represents an n-dimensional identity matrix.

The model is called a first order spatial autoregression since it represents a spatial analogy to the first order autoregressive model from time series analysis, that is, $y_t = \lambda y_{t-1} + \varepsilon_t$, where total reliance is on the past period observations to explain the variation in y_t. This model says that the levels of the dependent variable y depend on the levels of y in the neighboring regions. It is thus a formulation of the idea of a spatial spillovers. The FAR process may not be useful alone in empirical econometrics but it is used to model possible spatial correlations in disturbances of a regression equation.

We can extend the first-order spatial autoregressive model to include a matrix X of explanatory variables such as those used in traditional regression models. Anselin (1988) provides a maximum likelihood method for estimating the parameters of this model that he labels a "mixed regressive-spatial autoregressive model". We will refer to this model as the spatial autoregressive model (SAR). The SAR model takes the form:

$$\begin{aligned} y &= \lambda W y + X\beta + u \\ u &\sim N\left(0, \sigma^2 I_n\right) \end{aligned} \tag{1}$$

where y contains an $(n \times 1)$ vector of dependent variables, X represents the usual $(n \times k)$ data matrix containing explanatory variables and W is a known spatial weights matrix, usually a first-order contiguity matrix and where u is assumed to be the classical error term. The parameter λ is a coefficient on the spatially lagged dependent variable (Wy) and the parameters β reflect the influence of the explanatory variables on variation upon the dependent variable y. The model is termed a mixed regressive–spatial autoregressive model because it combines the standard regression model with a spatially lagged dependent variable, reminiscent of the lagged dependent variable model from time-series analysis. If there is no spatial dependence, and y does not depend on neighboring y values, then $\lambda = 0$. Theoretically, this model suggests some sort of diffusion/adoption process whereby our dependent variable is significantly influenced by its neighbors.

The term λWy measures the potential spillovers effect that occurs in the outcome variable if, this outcome is influenced by other unit's outcomes, where the location or distance to other observations is a factor for this spillovers. In other words, the neighbors for each observation have greater (or in some cases less) influence on what happens to that observation, independent of the other explanatory variables (X). The magnitude and impact of the spatial spillovers depend on the specification of the weights matrix.

Notice that λWy it is well defined since the diagonal elements of W are zero, which implies that we do not have the circular specification that y_j on the left hand side is influenced by the same y_j on the right hand side. Clearly we would not want to run ordinary least squares (OLS) on this model, since the presence of y on both sides of the equation means that we have a correlation problem between errors and regressors, and the resulting estimates will be biased and inconsistent. But we can easily obtain the reduced form as

$$\begin{aligned} y &= \lambda Wy + X\beta + u \qquad (2)\\ (I - \lambda W)\, y &= X\beta + u \\ y &= (I - \lambda W)^{-1} X\beta + (I - \lambda W)^{-1} u \end{aligned}$$

where $(I - \lambda W)^{-1}$ is the spatial multiplier (assuming that the inverse exists). We should mention the existence of some potential problems here. First, the new error term $u^* = (I - \lambda W)^{-1} u$ is no longer homoskedastic. Second, and probably more fundamentally, the model is no longer linear-in parameters because of the new unknown parameter λ.

The choice of the weights matrix could have a substantial impact on model interpretation. If we consider $(I - \lambda W)^{-1}$ and if $|\lambda| < 1$, that the inverse matrix can be expanded in a power series as

$$(I - \lambda W)^{-1} = I + \lambda W + \lambda^2 W^2 + \lambda^3 W^3 + \ldots$$

This expansion of the spatial multiplier illustrates the impact on the second-order, third-order, and higher-order neighbors. Note that, similar to the AR(1) coefficient in time series analysis, the dependent variable is on both sides of the equation. Interpretation of these models, however, is considerably more complex than AR(1) models. In AR(1) models, time moves on only in two directions: backwards and forwards. Spatial models are multidirectional, meaning that space, by its nature, moves in multiple directions.

4.2 SEM: The Spatial Error Model

In the spatial lag model, spatial dependence was assumed to be present in the dependent variable. Now we turn attention to the spatial errors model, where we incorporate spatial effects through the error (disturbance) term (not through the dependent variable). Following LeSage and Pace (2009);

$$\begin{aligned} y &= X\beta + u \\ u &= \rho W u + v \\ v &\sim N\left(0, \sigma^2 I_n\right) \end{aligned} \tag{3}$$

where y contains an $n \times 1$ vector of dependent variables, X represents the usual $n \times k$ data matrix containing explanatory variables, W is the spatial weights matrix, ρ represents the spatial error parameter, v is the vector of normally-distributed residuals with $E(v) = 0$, $E(vv') = \sigma^2 I$, and u represents the disturbance which is no longer iid normal in our case. The parameters reflect the influence of the explanatory variables on variation upon the dependent variable y. If there is no spatial correlation between the errors, then $\rho = 0$, and the model is reduced to the standard non-spatial linear regression model.

Solving the error specification for u we find

$$(I - \rho W)\, u = v \implies u = (I - \rho W)^{-1} v$$

and the model can be written as

$$y = X\beta + (I - \rho W)^{-1} v$$

This is conceptually simpler than the SAR model because the only problems occurring here are heteroskedasticity and non-linearity in ρ.

Theoretically this model is less compelling with respect to what it tells us about spatial processes. A spatial lag in the error term addresses missing variables with spatially distinct effects, and it is also employed to counter heterogeneity in observational units and sampling patterns.

4.3 Spatial Durbin Model

In time series analysis, independent variables are sometimes lagged by one or two time periods to see whether the previous values of those variables have significant effects or not. Similarly, independent variables can be spatially lagged. However, since space is multidirectional, the value of these spatially lagged variables is dependent on the weights matrix. A variant of the spatial lag model that spatially

lags all independent variables is known as the Spatial Durbin Model (SDM) (LeSage and Pace 2009) and is given by

$$y = \lambda Wy + X\beta + WX\theta + u \tag{4}$$
$$u \sim N\left(0, \sigma^2 I_n\right).$$

This model, for example, allows for characteristics that determine commuting times (variables contained in the matrix X) from neighboring regions to exert an influence on commuting times of region i. This is accomplished by entering an average of the explanatory variables from the neighboring regions, created using the matrix product WX. Apart from potential problems of multicollinearity, this model poses no major problems. The use of this model can potentially remove omitted variable bias, as discussed in detail by LeSage and Pace (2009).

Spatial spillovers are of main interest in regional science. Contrary to standard econometric models which restrict spillovers to be zero, a valuable aspect of spatial econometric models is that the magnitude and significance of this kind of spillovers can be empirically assessed. This is one of the reasons why spatial econometric methods are frequently used in regional science and have also seen increasing use in other scientific fields.

Until recently, empirical studies used the coefficient estimates of a spatial econometric model to test the hypothesis as to whether or not spatial spillovers effects exist. LeSage and Pace (2009) showed that a partial derivative interpretation of the impact from changes in the variables represents a more valid tool for testing this hypothesis. These spatial spillovers arise as a result of impacts passing through the neighboring regions and moving back to the region itself. The magnitude of this type of feedback will depend upon: (a) the position of the region in space, (b) the degree of connectivity among regions governed by the weight matrix W used in the model, (c) the parameter ρ measuring the strength of spatial dependence, and (d) the magnitude of the coefficient estimates for β and θ.

For example, if we rewrite the model in (4) as in (5), this will be useful for examining the partial derivatives of y with respect to a change in the rth variable x_r from X, as follows

$$y = (I_n - \lambda W)^{-1}\left[X\beta + WX\theta + u\right] \tag{5}$$
$$\frac{\partial y}{\partial x_r} = (I_n - \lambda W)^{-1}\left[I_n\beta_r + W\theta_r\right]$$

The partial derivatives are given by an $(n \times n)$ matrix, not by the typical scalar expression β_r from OLS. The matrix arises because a change in a single observation x_{ir} can influence all observations of the vector $y_j, j = 1, \ldots, n$. Considering changes in each of the x_{ir}, $i = 1, \ldots, n$ observations and the associated $(n \times 1)$ vectors of y-responses gives rise to the $(n \times n)$ matrix of partial derivatives. The own-region or direct effects are captured by the own-partial derivative $\partial y_i / \partial x_{ir}$ which are given by the elements on the diagonal of the matrix in (5). The cross-

partial derivatives $\partial y_j / \partial x_{ir}$, $j \neq i$ reflect indirect or spillover effects, and these are located on the off-diagonal elements of the matrix in (5).

We can summarize these differentiated effects by their mean. For example, the mean of $\partial y_j / \partial x_{ir}$, will define the average total effect of a unit change in x_r. Also, we can produce a decomposition of the average total effect of a unit change in all cells of x_r into a direct and indirect components. The average direct effect of a unit change in x_{ir} on y_i is given by the mean of the main diagonal of the matrix in (5). This direct effect is somewhat different from β_r because it also allows for the fact that a change in x_{ir} affects y_i, which in turn affects y_j, $(j \neq i)$ and so on, cascading through all areas and coming back to produce an additional (feedback) effect on y_i. The difference between the total effect and the direct effect is the average indirect effect of a variable. This is equal to the mean of the off-diagonal cells of the matrix (5).

4.4 The General Spatial Model

Spatial dependence can be modeled simultaneously, both in the dependent variable (SAR) and in the disturbances (SEM). There are two such general models: the spatial autocorrelation model (SAC), and the spatial autoregressive moving average model (SARMA), defined as:

$$\begin{aligned} y &= \lambda W_1 y + X\beta + u \\ u &= \rho W_2 u + v \end{aligned} \tag{6}$$

$$\begin{aligned} y &= \lambda W_1 y + X\beta + u \\ u &= \rho W_2 \varepsilon + \varepsilon \end{aligned} \tag{7}$$

where v is classical error term and both W_1 and W_2 are spatial weights matrices. One motivation for this is as follows. Suppose we have estimated a SAR model. We then test the residuals for spatial autocorrelation using (say) Moran's test. If we cannot reject the hypothesis that the residuals are (still) spatially autocorrelated, then this model, which allows for both sources, may be appropriate. A problem is the choice of forms (structure) for W_1 and W_2. Theory provides no guides, and if one takes $W_1 = W_2 = W$ then we can run into identification problems. In some cases, for example, if disturbance structure involves higher-order spatial dependence, then, a second-order spatial contiguity matrix can be used for W_2 that corresponds to a first-order contiguity matrix W_1.

4.5 Some Applications

The spatial econometric models presented in the previous sections have been used in several applications mostly within economics and regional science. In the context of the European Union, we should emphasize several empirical works in the particular areas of urban economics, economic growth and productivity, and studies dealing with agglomeration and externalities (spillovers). In this section, we will provide a brief survey of some of the results obtained in these particular areas. In the literature one may find several papers reviewing the application of spatial regression models, e.g., Anselin (2003, 2010), Abreu et al. (2005), Fingleton (2003), Getis et al. (2004), Gorter et al. (2005), LeSage and Fischer (2008), among others. Almost invariably these specifications are elaborations of mainstream theory incorporating externalities in the form of a spatial spillovers, being characterized by the presence of the standard component of the spatial econometric model (the spatial lag), which can be considered as the *sine qua non* condition of spatial econometrics.

Baumont et al. (2002) estimate the convergence of European regions and emphasize geographic spillovers in regional economic growth phenomena. In a sample of 138 European regions over the 1980–1995 period, they show that the unconditional β-convergence model is misspecified due to spatially autocorrelated errors. By using spatial econometric methods (the spatial autoregressive model, the spatial cross-regressive model and the spatial error model) and a distance based weight matrix, they estimate an alternative specification which takes into account the spatial autocorrelation detected and leads to reliable statistical inference. This specification allows to highlight a geographic spillovers effect: the mean growth rate of a region is positively influenced by those of neighboring regions. Other papers with related studies are Basile (2009), Dall'erba and LeGallo (2008), LeGallo and Baumont (2006), Rey and LeGallo (2009).

Paas and Schlitte (2007), offer empirical insights for the development of disparities in regional per capita GDP and convergence processes in the enlarged EU. A cross-section of 861 regions is analyzed for the period from 1995 to 2003. They conduct a formal β-convergence analysis, taking into account the effects of spatial dependence and controlling for national effects. The analyses show that poorer regions mainly situated in the European periphery have tended to grow faster than the relatively rich regions in the centre of Europe. Furthermore, the authors find that spatial growth spillovers loose relevance when crossing a national border. Thus, border constraints are still of great relevance for the intensity of economic cross-border integration within the EU.

Niebuhr (2003) focuses on the spatial structure of regional unemployment disparities. Regions are tightly linked by migration, commuting and interregional trade. These types of spatial interaction are exposed to the frictional effects of distance, possibly causing the spatial dependence of regional labour market conditions. The spatial association of regional unemployment is analyzed for a sample of European

countries between 1986 and 2000 by measures of spatial autocorrelation and spatial regression models. The results indicate that there is a significant degree of spatial dependence among regional labour markets in Europe. Regions marked by high unemployment, as well as areas characterized by low unemployment, tend to cluster across space. These findings suggest that different forms of spatial interaction affect the evolution of regional unemployment in Europe.

Moreno et al. (2005) explore the spatial distribution of innovative activity and the role of technological spillovers in the process of knowledge creation across 138 regions of 17 countries in Europe (the 15 members of the European Union plus Switzerland and Norway). They proceed to an exploratory spatial data analysis of the dissemination of innovative activity in Europe and present some global and local indicators for spatial association, summarizing the presence of a dependence process in the distribution of innovative activity for different periods and sectors. They also attempt to model the behavior of innovative activity at the regional level on the basis of a knowledge production function. Econometric results (spatial cross-section models) indicate that internal factors seem to be relevant for innovative activities (like R&D expenditure, economic performance, agglomeration economies), and the results show also that the production of knowledge by European regions seems to be affected by spatial spillovers due to innovative activity performed in other regions as well.

Baldacci et al. (2011) use spatial econometrics techniques (different weighting matrices, Moran's I statistics, spatial autoregressive and spatial error models) to explore spillovers in the sovereign bond market for 24 emerging economies during 1995–2010. The paper extends the previous literature focusing on spillovers effects from advanced to emerging economies by analyzing transmission of shocks across emerging markets. After controlling for the impact of global factors, they find strong evidence of spillovers from both sovereign spreads and macroeconomic fundamentals in the neighboring emerging economies. In addition to the geographical proximity, the channels of spatial transmission include trade and financial linkages. More results on spillovers effects can be find in the Vega and Elhorst (2013), and Anselin (2010).

Head and Mayer (2004) examine empirical strategies that have been used to evaluate the importance of agglomeration and trade models. This theoretical approach, known as the "New Economic Geography", emphasizes the interaction between transport costs and firm-level scale economies as a source of agglomeration. Fingleton (2000), Gorter et al. (2005) and Redding (2010), review the existing empirical literature on the predictions of new economic geography models for the distribution of income and production across space. The discussion highlights connections with other research in regional and urban economics, identification issues, potential alternative explanations and possible areas for further research.

4.6 Estimation

The main steps that a researcher should follow when dealing with a spatial model are briefly presented below:

1. Map the data and choose a neighborhood criterion;
2. Create a spatial weights matrix;
3. Run a statistical test to examine spatial autocorrelation (Moran's I test);
4. Run an OLS regression, check residuals and determine what type of spatial model to use;
5. Run a spatial regression model;
6. Analyze the equilibrium and the feedback implications of the estimated spatial model for the dependent variable.

In a classical time-series model with a lagged term, y_{t-1}, on the right hand side of the equation, the presence of the temporal lag y_{t-1} does not create problems for estimation with OLS, if there is no serial correlation in the residuals of the regression model (the model is correctly specified). However, in the spatial case, if y_{t1} is predetermined at time $t1$, the spatial lag of y is simultaneous and based on y itself. This simultaneity creates problems when estimating the spatially lagged y model. The estimation of the specification remains unbiased, but is no longer efficient, and the classical estimators for standard errors will be biased (Anselin 1988). The two main approaches to the estimation of spatial models are based on the maximum likelihood principle and the generalized method of moments. More robust methods that account for heteroskedasticity use either two-stage least squares/instrumental variables (Kelejian and Prucha 1999; Bivand 2010) or generalized method of moments (Kelejian and Prucha 2007; Piras 2010).

4.6.1 ML Estimation of the SAR Model

We recall that the SAR model given as

$$y = \lambda Wy + X\beta + u$$

can be written in the following way

$$\begin{aligned}(I - \lambda W)\, y &= X\beta + u \\ Ay &= X\beta + u\end{aligned}$$

where $u \sim N\left(0, \sigma^2 I\right)$, $A = (I - \lambda W)$ and W is a row-standardized spatial weights matrix. The model is usually estimated by ML method. The log-likelihood function is given by

$$\ln L\left(\beta, \lambda, \sigma\right) = -\left(\tfrac{n}{2}\right) \ln \pi - \left(\tfrac{n}{2}\right) \ln \sigma^2 + \ln \|A\| - \left(\tfrac{1}{2\sigma^2}\right) (Ay - X\beta)' (Ay - X\beta)$$

where $\|A\|$ is the determinant of A. Anselin (1988), suggests a way to do the estimation. If we concentrate first on β, it is easy to show that the ML estimator is given by

$$\begin{aligned} b &= \left(X'X\right)^{-1} X'Ay \\ &= \left(X'X\right)^{-1} X'y - \lambda \left(X'X\right)^{-1} X'Wy = b_0 - \lambda b_L \end{aligned}$$

where $b_0 = (X'X)^{-1} X'y$ and $b_1 = (X'X)^{-1} X'Wy$. Some close verification shows that b_0 is the coefficient vector from the OLS regression of y on X, while b_L is from the OLS regression of Wy on X. So if λ is known, we can compute the ML estimate of β. Next, we write the residuals of these two OLS regressions as

$$\begin{aligned} e_0 &= y - Xb_0 \\ e_L &= Wy - Xb_L \end{aligned}$$

and it can be shown that the ML estimate of σ^2 is given by

$$s^2 = \left(\frac{1}{n}\right)(e_0 - \lambda e_L)'(e_0 - \lambda e_L)$$

so once again we could estimate σ^2 if λ were known.

We use all this information to write down a version of the log-likelihood function only in terms of λ and we obtain the concentrated log-likelihood, $\ln L^*$.This is given by

$$\ln L^* = C - \left(\frac{n}{2}\right) \ln \left[\left(\frac{1}{n}\right)(e_0 - \lambda e_L)'(e_0 - \lambda e_L)\right] + \ln \|A\|$$

where C contains only known parameters. We proceed now with the maximization of $\ln L^*$ with respect to λ and obtain the ML estimate of this parameter, and work backwards.

Summarizing, the estimation steps are the following:

1. Regress y on X : we obtain b_0. Compute the residual $e_0 = y - Xb_0$;
2. Regress Wy on X : we obtain b_L. Compute the residual $e_L = Wy - Xb_L$;
3. Find the λ that maximizes the concentrated log-likelihood function. Call it $\hat{\lambda}$;
4. Given $\hat{\lambda}$, compute $b = b_0 - \hat{\lambda} b_L$ and

$$s^2 = (1/n)\left(e_0 - \hat{\lambda} e_L\right)'\left(e_0 - \hat{\lambda} e_L\right)$$

We can observe that the first two steps are simply OLS linear estimation problems and the third one is a one-parameter nonlinear optimization problem that can be solved with the adequate numerical algorithm. One problem which occurs here, due

to the stepwise nature of the estimation process, is that we don't get the estimates of the (joint) covariance matrix of all the estimated parameters. However, since they are maximum-likelihood estimates, we know that they are asymptotically efficient, that is, for large samples the covariance matrix attains the Cramer–Rao lower bound, given by

$$-E\left(\frac{\partial^2 \ln L}{\partial\theta\partial\theta'}\right)^{-1}$$

where $\theta = \left(\beta, \lambda, \sigma^2\right)$. This in turn can be estimated by the numerical Hessian of the log likelihood.

4.6.2 ML Estimation of the SEM Model

Let consider now the cross-sectional SEM model

$$\begin{aligned} y &= X\beta + u \\ u &= \rho W u + v \end{aligned}$$

where v is the error term, and which can be written as

$$\begin{aligned} y &= X\beta + u \\ Bu &= v, \quad B = (I - \rho W). \end{aligned}$$

The estimation of this model is more complicated comparing with the SAR model. The log-likelihood function is given by

$$\ln L = -\left(\frac{n}{2}\right)\ln\pi - \left(\frac{n}{2}\right)\ln\sigma^2 + \ln\|B\| - \left(\frac{1}{2\sigma^2}\right)(y - X\beta)'\, B'B\,(y - X\beta)$$

and, as previously, Bv is heteroskedastic. We can estimate β using GLS (generalized least squares), and then, an estimate of σ^2 is similar to the SAR case. The concentrated log-likelihood is defined by

$$\ln L^* = C - \left(\frac{n}{2}\right)\ln\left[\left(\frac{1}{n}\right)e'B'Be\right] + \ln\|I - \rho W\|$$

where $e = y - Xb_{GLS}$. The main problem here is that b_{GLS} itself depends on ρ (unlike the SAR case).

Anselin (2010) suggests an iterative procedure, which basically consists of the following steps:

1. Regress y on X. Call the coefficient estimate b_{OLS} and compute the residual vector $e = y - Xb_{OLS}$;
2. Use this e in the concentrated log-likelihood, and optimize to find $\hat{\rho}$;
3. Use $\hat{\rho}$ to compute the GLS estimator b_{GLS} and then a new residual vector $e = y - Xb_{GLS}$;
4. First time or if the residuals have not converged: go back to step 2 and reestimate ρ. Otherwise: go to step 5;
5. At this point we have a converged estimate of ρ (say, $\hat{\rho}$ and the associated residual vector e, and a GLS estimator of β. We can now estimate σ^2 by $(1/n)\, e' B' Be$.

5 Software

Most standard statistics packages do not contain estimation routines for spatial econometric models but there are several software packages that can perform spatial econometric-related tests and estimate spatial econometric models. For example, GeoDa and OpenGeoDa software (Anselin 2010), are available with no cost at the GeoDa Center. These are used for exploratory data analysis and for preliminary tests of spatial correlation. LeSage (2010) has published a MATLAB-based econometrics toolbox, which can estimate many spatial econometric models, both using maximum likelihood and Bayesian methods. Documentation for the toolbox includes a special section for the spatial econometrics commands as well as learning materials related to spatial econometrics (LeSage 1998). A similar toolbox was developed for STATA, containing regression diagnostics, maximum likelihood and GMM estimation (Pisati 2001, 2008). The majority of the spatial econometric models in the literature can be estimated using packages in the open-source R statistical package, see for example Bivand (2010), Bivand et al. (2008), Piras (2010).

References

Abreu, M., de Groot, H., & Florax, R. (2005). Space and growth: A survey of empirical evidence and methods. *Région et Développement, 21*, 12–43.

Anselin, L. (1988). *Spatial econometrics: Methods and models*. Boston: Kluwer Academic.

Anselin, L. (2003). Spatial externalities, spatial multipliers, and spatial econometrics. *International Regional Science Review, 26*(2), 153–166.

Anselin, L. (2006). Spatial econometrics. In T. Mills, & K. Patterson, (Eds.), *Palgrave handbook of econometrics*. Hampshire and New York, NY: Palgrave Macmillan.

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in Regional Science, 89*(1), 3–25.

Arbia, G. (2006). *Spatial econometrics: Statistical foundations and applications to regional convergence (advances in spatial science)*. Berlin: Springer.

Baldacci, E., Dell'Erba, S., & Poghosyan, T. (2011). *Spatial spillovers in emerging market spreads*. IMF working paper, 11/221.

Basile, R. (2009). Productivity polarization across regions in Europe: The role of nonlinearities and spatial dependence. *International Regional Science Review, 32*(1), 92–115.

Baumont, C., Ertur, C., & Le Gallo, J. (2002). The European regional convergence process, 1980–1995: Do spatial regimes and spatial dependence matter? *Econometrics* 0207002, EconWPA.

Bivand, R. (2010). *Spatial dependence: Weighting schemes, statistics and models*. R package version 0.5-4. http://CRAN.R-project.org/ package =spdep.

Bivand, R. S., Pebesma, E. J., & Gómez-Rubio, V. (2008). *Applied spatial data analysis with R, Use R!* Secaucus, NJ: Springer.

Dall'erba, S., & LeGallo, J. (2008). Regional convergence and the impact of European structural funds over 1989–1999: A spatial econometric analysis. *Papers in Regional Science, 87*(2), 219–244.

Fingleton, B. (2000). Spatial econometrics, economic geography, dynamics and equilibrium: A 'third way'? *Environment and Planning A, 32*(8), 1481–1498.

Fingleton, B. (2003). Externalities, economic geography, and spatial econometrics: Conceptual and modeling developments. *International Regional Science Review, 26*(2), 197–207.

Fingleton, B. (2004). Some alternative geo-economics for Europe's regions. *Journal of Economic Geography, 4*(4), 389–420.

Getis, A., Mur, J., & Zoller, H. (Eds.) (2004). *Spatial econometrics and spatial statistics*. Basingstoke: Palgrave Macmillan.

Gorter, J., van der Horst, A., Brakman, S., Garretsen, H. F. L., & Schram, M. (2005). *New economic geography, empirics, and regional policy*. CPB Special Publication 56, CPB Netherlands Bureau for Economic Policy Analysis.

Griffith, D. A., & Paelinck, J. H. P. (2011). *Non-standard spatial statistics and spatial econometrics. Series: Advances in geographic information science*. Berlin: Springer.

Head, K., & Mayer, T. (2004). *The empirics of agglomeration and trade*. Sciences Po Publications info:hdl:2441/10191, Sciences Po.

Hordijk, L., & Paelinck, J. H. P. (1976). Spatial econometrics, some contributions. In J. Paelinck (Ed.), *Actes des Tables Rondes 1974–1975 de l'Association de Science Régionale de Langue Française* (pp 125–143). Rotterdam.

Kelejian, H. H., & Prucha, I. R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review, 40*(2), 509–533.

Kelejian, H. H., & Prucha, I. R. (2007). HAC estimation in a spatial framework. *Journal of Econometrics, 140*, 131–154.

LeGallo, C., & Baumont, C. (2006). The European regional convergence process, 1980–1995: Do spatial regimes and spatial dependence matter? *International Regional Science Review, 29*(1), 3–34.

LeSage, J. P. (1998). *Spatial econometrics*. http://www.spatialeconometrics.com/html/wbook.pdf.

LeSage, J. P. (2010). *Econometrics toolbox*. http://www.spatialeconometrics.com/.

LeSage, J. P., & Fischer, M.M. (2008). Spatial growth regressions: Model specification, estimation and interpretation. *Spatial Economics Analysis 3*(3), 275–304.

LeSage, J. P., & Pace, R.K. (2009). *Introduction to spatial econometrics*. Boca Raton: Taylor and Francis/CRC Press.

Moreno, R., Paci, R., & Usai, S. (2005). Spatial spillovers and innovation activity in European regions. *Environment and Planning A, 37*(10), 1793–1812.

Niebuhr, A. (2003). Spatial interaction and regional unemployment in Europe. *European Journal of Spatial Development, 5*, 1650–9544.

Paas, T., & Schlitte, F. (2007). *Regional income inequality and convergence processes in the EU-25*. HWWI research papers, 1–11, Hamburg Institute of International Economics.

Paelinck, J. H. P., & Klaassen, L. H. (1979). *Spatial econometrics*. Saxon House Farnborough.

Piras, G. (2010). *Sphet: Spatial models with heteroskedastic innovations*. R package version 0.1-22. http://CRAN.R-project.org/package=sphet.

Pisati, M. (2001). Tools for spatial data analysis. *Stata Technical Bulletin, 60*, 21–37.

Pisati, M. (2008). *SPMAP: Stata module to visualize spatial data*. Stata Help Files. http://econpapers.repec.org/software/bocbocode/s456812.htm.

Redding, S. J. (2010) The empirics of new economic geography. *Journal of Regional Science, Wiley Blackwell, 50*(1), 297–311.

Rey, S. J., & LeGallo, J. (2009). Spatial analysis of economic convergence. In T.C. Mills & K. Patterson (Eds.), *Palgrave handbook of econometrics. Applied econometrics* (Vol. II, pp. 1251–1293).

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography, 46*(2), 234–240.

Vega, S. H., & Elhorst, J. P. (2013). *On spatial econometric models, spillovers effects*. W. Working paper, University of Groningen.

Semiparametric Spatial Autoregressive Geoadditive Models

Roberto Basile, Saime Kayam, Román Mínguez, Jose María Montero, and Jesús Mur

Abstract Modeling regional economic dynamics requires the adoption of complex econometric tools, which allow us to deal with some important methodological issues, such as spatial dependence, spatial heterogeneity and nonlinearities. Recent developments in the spatial econometrics literature have provided some instruments (such as Spatial Autoregressive Semiparametric Geoadditive Models), which address these issues simultaneously and, therefore, are of great use for practitioners. In this paper we describe these methodological contributions and present some applications of these methodologies in the fields of regional science and economic geography.

1 Introduction

Modeling regional economic dynamics requires the adoption of complex econometric tools, which allow us to deal with some important methodological issues, such as spatial dependence, spatial heterogeneity and nonlinearities. Regional and urban growth theories provide good examples to illustrate these issues. Recent develop-

R. Basile (✉)
Department of Economics, Second University of Naples, Corso Gran Priorato di Malta, 81043 Capua (CE), Italy
e-mail: roberto.basile@unina2.it

S. Kayam
Department of Management Engineering, Istanbul Technical University, Suleyman Seba C. No. 90, Macka, Istanbul, Turkey
e-mail: kayams@itu.edu.tr

R. Mínguez • J.M. Montero
Statistics Department, University of Castilla-La Mancha, Cuenca, Spain
e-mail: roman.minguez@uclm.es; jose.mlorenzo@uclm.es

J. Mur
Department of Economic Analysis, University of Zaragoza, Pedro Cerbuna 12, 50009 Zaragoza, Spain
e-mail: jmur@unizar.es

P. Commendatore et al. (eds.), *Complexity and Geographical Economics*,
Dynamic Modeling and Econometrics in Economics and Finance 19,
DOI 10.1007/978-3-319-12805-4_4

ments in economic growth theory have proposed extensions of multi-region growth models (see, i.a., Ertur and Koch 2007) that include technological interdependence across regions, in order to consider neighborhood effects (*spatial dependence*) in growth and convergence processes. Relaxing the strong homogeneity assumptions on the cross-region growth process and starting from the more realistic hypothesis that different economies should be described by distinct production functions, other contributions to the economic growth theory have cited the issue of *nonlinearities* to explain club convergence (see, i.a., Durlauf et al. 2005). Nonlinearities and spatial dependence have also been addressed in the literature on the effect of agglomeration economies on local (urban) economic growth. For example, a hump-shaped relationship between economic density and local productivity (or employment) growth has been highlighted by Basile et al. (2013): the positive effect of agglomeration externalities fades as the density of economic activities reaches some threshold value, after which negative effects due to congestion costs prevail. In modeling the effects of agglomeration economies on local economic growth, it is important to recognize that externalities may overcome the administrative boundaries of the regions, thus generating spatial dependence. Finally, as pointed out by Krugman (1993), the marked unevenness of local economic development can be partly justified on the basis that space is not uniform (*spatial heterogeneity*): "first nature" characteristics of local areas (i.e. unobserved spatial heterogeneity) must be carefully controlled for when specifying a local economic growth model, especially when these unobservables are potential sources of endogeneity.

Many other examples in regional science and economic geography may be used to discuss the issues of nonlinearities, spatial dependence and spatial heterogeneity. Here, we want to point out that applied econometric studies rarely tackle all these issues simultaneously. Most studies focus on one of these matters, disregarding the interdependence between them. For example, some scholars focus on detecting the presence of spatial dependence while assuming a linear functional form for the data generating process (Rey and Gallo 2009; Paci and Usai 2008), others only try to assess the existence of nonlinear effects (Ertur and Gallo 2009), some others only control for spatial heterogeneity in a panel framework (Henderson 1997). However, nonlinearities, spatial dependence and spatial heterogeneity are not orthogonal issues and disregarding one of them may generate some biases. For example, McMillen (2003) shows that specification tests may indicate spatial autocorrelation when functional form misspecification is actually the only problem with the model. Thus, incorrect functional forms and omitted variables that are correlated over space produce spurious spatial autocorrelation. Basile and Gress (2005) have also provided evidence of a trade-off between spatial autocorrelation and nonlinearities: the value of the spatial auto-correlation parameter is lower when possible nonlinearities are taken into account.

Recent developments in the spatial econometrics literature have provided some instruments (such as *Spatial Autoregressive Semiparametric Geoadditive Models*), which address the three issues simultaneously and, therefore, are of great use for practitioners (Gress 2004; Basile 2009; Su and Jin 2010; Su 2012; Mínguez et al. 2012; Montero et al. 2012). In this paper we describe some of these methodological

contributions and present some applications in the field of regional science and economic geography. We start by briefly reviewing the broad literature on parametric spatial econometric models and raising some critical issues concerning these models (Sect. 2). Then, we introduce semiparametric geoadditive models and describe their potential (Sect. 3). Specifically, we describe a control function approach to estimate a spatial lag semiparametric geoadditive model. In Sect. 4 we discuss an alternative way to specify (spatial lag) semiparametric geoadditive models as mixed models. In Sect. 5 we present selected empirical works using these models in the field of regional science and economic geography in order to show the wide scope for applications of this approach. In particular, we present an application to regional growth. Concluding remarks are reported in Sect. 6.

2 Parametric Spatial Econometric Models

Regional economics works within a spatial realm, and that means heterogeneity (i.e., heteroskedasticity and spatial instability of parameters) and interdependence (i.e., spatial dependence). Spatial econometrics obviously deals with these two topics, predominantly under a parametric approach. This section discusses some of the key aspects of this strategy of building regional models and highlights some of its limits.

Let y_i be the response variable computed for spatial unit i and $\{x_{k,i}; k = 1, 2, \ldots K\}$ a set of K explanatory variables. For example, y_i may measure the productivity growth rate of region i computed for a sufficiently long time span, whereas the x variables may include different indices of human capital and investment effort computed for the same time span. A key element in current spatial econometrics is the Spatial Durbin Model (SDM):

$$y_i = \beta_0 + \sum_{k=1}^{K} \beta_k x_{k,i} + \varrho \sum_{j=1}^{n} w_{ij} y_j + \sum_{k=1}^{K} \theta_k \sum_{j=1}^{n} w_{ij} x_{k,j} + \varepsilon_i \quad (1)$$

where ε_i is a white noise normally distributed error term, $\varepsilon_i \sim iid\mathcal{N}(0; \sigma_\varepsilon^2)$. This equation includes a sequence of (spatial linear) weights for the purpose of measuring the influence received by region i from region j: $\{w_{ij}; j = 1, 2, \ldots n\}$. Using these weights we can obtain spatial lags of the variables of interest. Using matrix notation, we get:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varrho \mathbf{W}_n \mathbf{y} + \mathbf{W}_n \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \quad (2)$$

where $\mathbf{y}$ is the $(n \times 1)$ vector of observations of the explained variable on the n regions, $\mathbf{X}$ is a $(n \times K)$ matrix of observations of the explanatory variables on the same regions and $\boldsymbol{\varepsilon}$ is a $(n \times 1)$ vector of error terms. Moreover, $\mathbf{W}_n$ is a $(n \times n)$ weighting matrix, assumed to be the same in the spatial lag of $\mathbf{y}$ and in the lags of the xs variables (the assumption can be relaxed).

Assuming $\boldsymbol{\theta} = \mathbf{0}$ in Eq. (2), a Spatial Autoregressive (SAR) Model is obtained; a Spatial Lag of $\mathbf{X}$ Model (SLX) results from the restriction that $\varrho = 0$; and a Spatial Error Model (SEM) appears under the assumption that $\boldsymbol{\theta} - \varrho\boldsymbol{\beta} = \mathbf{0}$; the pure non-spatial model would occur if $\varrho = \boldsymbol{\theta} = 0$. The puzzle may also include a SDM or a SAR model with spatially autocorrelated errors, $\boldsymbol{\varepsilon} = \varphi\mathbf{W}_n\boldsymbol{\varepsilon} + \boldsymbol{\eta}$, $\boldsymbol{\eta}$ being a white noise random vector. Other specifications can be found in LeSage and Pace (2009) and Elhorst (2010).

The term $\mathbf{W}_n\mathbf{y}$ that appears in the right-hand side (*rhs* from now on) of (2) is correlated with the error term, $Cov\left[\mathbf{W}_n\mathbf{y}; \boldsymbol{\varepsilon}\right] \neq \mathbf{0}$, so that ordinary least squares (OLS) estimates are biased and inconsistent. Consistent and efficient estimates can be obtained by maximum likelihood (ML) or quasi-maximum likelihood (QML) estimates, if the assumption of normality cannot be maintained (Lee 2004). Two-Stage Least Squares (2SLS) estimates adapt well to the case of (2) because higher orders of spatial lags of the x variables are natural candidates to be used as instrumental variables (Kelejian and Prucha 1997). A more efficient estimator is the method of moments estimator (MM) (Kelejian and Prucha 2001). Lee (2004) generalized the MM approach into a fully generalized method of moments (GMM) estimator for the case of the SDM model (2), while Liu et al. (2007) proposed a GMM estimator for a SDM with dependent structures in the error term. The GMM estimator may have, under general conditions, the same limiting distribution as the ML or QML estimators. Moreover, the GMM estimator allows the researcher to take into account any endogeneity problems in the *r.h.s.*, different from the spatial lag of $\mathbf{y}$.

The above-mentioned spatial econometric models allow for interdependence among regions. All of them correspond to a long-run equilibrium relation between the response variable and its covariates; time dynamics is ruled out. In response to an exogenous variation in an x variable (such as, e.g., an improvement in the human capital stock in Southern Italy's provinces), these spatial models return the expected impact on the dependent variable (e.g., the productivity growth rates) for the whole regional system in the steady state solution.[1] It is customary to distinguish between *local* and *global* spatial spillovers (Anselin 2003). The key is the existence of feedback effects in the equation. Model (2) contains feedback effects: a change in a regressor in region i impacts on the outcome of this region, on the outcome of its neighbors, on that of the neighbors of its neighbors and so on. The impact therefore is *global*. On the contrary, the multipliers obtained from a SLX model ($\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}_n\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$) are *local*, since the impact of the change dies just after its effect on the neighbors.

When spatial spillovers are global, we may distinguish between *direct* and *indirect* spatial effects. *Direct* effects measure the impact of a change in regressor k in region i on the outcome of the same region: $\frac{\partial y_i}{\partial x_{ki}}$, while *indirect* effects measure

[1] If the interest lies in the short-term adjustments, a spatial panel data model would be required instead (Elhorst 2012).

the impact of a change in regressor k in region j on the outcome of region i: $\frac{\partial y_i}{\partial x_{kj}}$. The problem with these effects is that, conditional on the model, they are specific to the pair of regions involved (i, j).[2] For this reason LeSage and Pace (2009) propose the use of average indicators. For example, in the SDM model of (2), the *average direct* effect of variable k is:

$$\overline{de_k} = \sum_{i=1}^{n} \frac{(\mathbf{I}_n - \varrho \mathbf{W}_n)^{-1} [\mathbf{I}_n(\beta_k + \theta_k \mathbf{W}_n)]_{ii}}{n}; \; k = 1, 2, \ldots., K \tag{3}$$

This is the mean value of the n *direct* effects (the sub-index ii means the ith element of the main diagonal of the square matrix $(\mathbf{I}_n - \varrho \mathbf{W}_n)^{-1} [\mathbf{I}_n(\beta_k + \theta_k \mathbf{W}_n)]$. In order to obtain the average *indirect* effect we must first define the average *total* effect:

$$\overline{te_k} = \sum\nolimits_{j=1}^{n} \sum\nolimits_{i=1}^{n} \frac{(\mathbf{I}_n - \varrho \mathbf{W}_n)^{-1} [\mathbf{I}_n(\beta_k + \theta_k \mathbf{W}_n)]_{ij}}{n^2}; \; k = 1, 2, \ldots., K \tag{4}$$

The *average total* effect measures the accumulated impact of a change in the x_k variable on the dependent variable, $\mathbf{y}$, located in any region. The difference between the two corresponds to the *average indirect* effect:

$$\overline{ie_k} = \overline{te_k} - \overline{de_k}; \; k = 1, 2, \ldots., K \tag{5}$$

and measures the importance of the spatial spillovers in the model.

A very important issue in this context concerns the choice of the weighting matrix whose purpose is to reflect the arrangement of the space. Uncertainty is a big problem here because the researcher must provide this matrix and, normally, he/she has a very limited knowledge about it. Of course, the problem of selecting the right $\mathbf{W}_n$ is reflected on the computation of spatial multipliers: a bad choice of the matrix invalidates subsequent analysis. A few rules can be provided to tackle this issue: $\mathbf{W}_n$ must be a square $(n \times n)$ matrix whose elements are, usually, non-negative and its main diagonal is comprised of zeroes (both restrictions can be relaxed). According to Kapoor et al. (2007), this matrix should be uniformly bounded in absolute value in order to assure convergence and, in order to avoid cases of isolation, the row sums should be uniformly bounded away from zero. It is typical to row-standardize the matrix before its use in estimation algorithms.

[2] Although classical spatial econometric models are based on the assumption of linearity and parameter homogeneity, they allow us to assess a form of heterogeneity, called "interdependence heterogeneity" (Ertur and Koch 2011): the magnitude of spatial direct and indirect partial effects is different among regions, since it depends upon the position of the regions in space, the degree of connectivity among regions, which is governed by the $\mathbf{W}_n$ matrix and the estimated model parameters.

Thus far we have sketched some of the key elements of spatial econometrics. Now, an important question remains: Why should econometric models take spatial effects into account? Fingleton and López-Bazo (2006) observe that if we are using spatial data it is reasonable to find spatial interaction:

> Put very simply, if we assume that firms are heterogeneous and always interacting with each other, then the fact that they are often located in different regions will cause regions to be heterogeneous and interdependent (p. 178).

This is what intuition tells us and the long list of mis-specification tests usually corroborates (Elhorst 2010). However, from a purely econometric perspective, we are forced to go a bit further because the indiscriminate use of the spatial lag of the endogenous variable can muffle the impact of other specification errors (Partridge et al. 2012). In this sense, LeSage and Pace (2009) point to misspecification problems due to (a) mismatches in the time structure of the equation; (b) omission of relevant explanatory variables from the *r.h.s.*; (c) the impact of omitted heterogeneity in the model. McMillen (2003) adds a fourth factor (d), the consequences of a wrong selection in the functional form of the equation. Kelejian and Robinson (2004) also observe that heteroskedasticity and spatial dependence produce similar signs and are very often confused. Similarly, spatial dependence tests react strongly under heterogeneity in the parameters (Mur et al. 2009).

These observations advocate for a more theory-driven approach. Examples in this direction are the regional production function with externalities in the rate of technological progress (López-Bazo et al. 2004; Ertur and Koch 2007, 2011), or the Verdoorn Law with knowledge spillovers proposed (Fingleton 2004). Spatial prices and spatial market competition models have also produced many papers in the same spirit (see, i.a., Pinkse et al. 2002; Holly et al. 2010). This practice is not typical, however. In fact, a kind of empiricism controls the process of building a model: first, a simple provisional equation is specified, then a large battery of misspecification tests is applied and, finally, spatial interaction mechanisms are introduced ad hoc as reaction to the tests. The consequence may be that, at the end, we lose perspective.

The observation that spatial lag terms may actually capture the effect of other specification errors may also suggest the need for a more flexible (semiparametric) approach, which relaxes the restrictive assumptions (linearity and parameter homogeneity) of the parametric approach. One step in this direction is the so-called Geographically Weighted Regression (*GWR*) model (Fotheringham et al. 2002; Lloyd 2011), which is a nonparametric (local linear) method to capture spatial parameter heterogeneity.[3]

[3]The *GWR* has been extended to cross-sectional models with spatial interaction terms by Pace and LeSage (2004) and Mur et al. (2009).

3 Semiparametric Geoadditive Models

In this section, we present a semiparametric framework, which allows us to relax the linearity assumption and simultaneously model spatial dependence and spatial heterogeneity. We start by introducing a general specification of the semiparametric geoadditive model without a spatial lag term (Sect. 3.1) and discussing technical issues concerning its estimation (Sects. 3.2 and 3.3). In Sect. 3.4 we extend this model by introducing the spatial lag of the dependent variable on the *rhs* so as to get a spatial autoregressive semiparametric geoadditive model (*SAR-Geo-AM*).

3.1 Model Specification

In Sect. 2 we pointed out that most of the spatial econometric literature focuses on the issue of spatial dependence. A strand of this literature also stresses the problem of spatial heterogeneity, that is, of spatial instability of the parameters (using the *GWR* method), while the issue of nonlinearity (i.e., the choice of the functional form) is strongly neglected. Obviously, nonlinearities might also be captured within a pure parametric framework by using a polynomial expansion, but this may lead to strong collinearity (unless orthogonal polynomials are used). Semiparametric methods represent a more satisfactory solution since they are more flexible than any parametric specification. By using a particular version of the semiparametric model that allows for additive components (Hastie and Tibshirani 1990), we are able to obtain graphical representation of the relationship between the response variable and the covariates. Additivity ensures that the effect of each predictor can be interpreted net of the effects of the other regressors, as in multiple linear regressions.[4]

The starting point may be a general form of the semiparametric additive model suitable for large cross-regional data:

$$y_i = \mathbf{X}_i^{*'}\boldsymbol{\beta}^* + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + f_4(x_{1i})\, l_i + \ldots + \varepsilon_i \quad (6)$$

$$\varepsilon_i \sim iid\mathcal{N}\left(0, \sigma_\varepsilon^2\right) \qquad i = 1, \ldots, n$$

where y_i is a continuous univariate response variable measuring, for example, the average annual productivity growth rate of region i. $\mathbf{X}_i^{*'}\boldsymbol{\beta}^*$ is the linear predictor for any strictly parametric component (including the intercept, all categorical covariates and eventually some continuous covariates), with $\boldsymbol{\beta}^*$ being a vector of

[4] Usually, a fully nonparametric model (i.e., a model where all terms are smoothly interacted with each other) cannot be applied to regional data since it would require a very large number of observations to overcome the curse of dimensionality. Additivity is therefore a valid compromise between flexibility and tractability.

fixed parameters. $f_k(.)$ are unknown smooth functions of univariate continuous covariates or bivariate interaction surfaces of continuous covariates capturing nonlinear effects of exogenous variables. Which of the explanatory variables enter the model parametrically or non-parametrically may depend on theoretical priors or can be suggested by the results of model specification tests (see, e.g., Kneib et al. 2009). $f_4(x_{1i})\, l_i$ is a varying coefficient term, where l_i is either a continuous or a binary covariate. For example, we may want to test whether the smooth effect of x_1 (e.g., population density) is different in the North and in the South. In this case l_i is a binary variable taking value one if region i belongs to the North and zero if it belongs to the South. Thus, if $l_i = 0$, the effect of x_1 is given by $f_1(x_{1i})$, whereas for $l_i = 1$, the effect is composed as the sum $f_1(x_{1i}) + f_4(x_{1i})$, and $f_4(x_{1i})$ can be interpreted as the deviation of x_1 for the North. Finally, ε_i are *iid* normally distributed random shocks.

Model (6) captures nonlinearities in the relationship between the response variable y_i and its covariates, but it does not take into account any spatial structure of the data. Removing *unobserved spatial patterns* is a primary task, especially when the researcher considers spatial unobservables as potential sources of endogeneity, that is, when there is a suspected correlation between unobserved and observed variables. This issue can be addressed within the semiparametric framework by incorporating the spatial location as an additional covariate in (6), thus generating what is known in the literature as the geoadditive model:[5]

$$\begin{aligned} y_i = {} & \mathbf{X}_i^{*'}\boldsymbol{\beta}^* + f_1(x_{1i}) + f_2(x_{1i}) + f_3(x_{3i}, x_{4i}) + f_4(x_{1i})\, l_i + \ldots \\ & + h(no_i, e_i) + \varepsilon_i \end{aligned} \tag{7}$$

The term $h(no_i, e_i)$ in Eq. (7) is a smooth spatial trend surface, i.e., a smooth interaction between latitude (*northing*) and longitude (*easting*). It allows us to control for unobserved spatial heterogeneity.

When the term $h(no_i, e_i)$ is interacted with one of the explanatory variables, (e.g., $h(no_i, e_i)\, x_{1i}$), it allows us to estimate spatially varying coefficients (like in the *GWR* model). For example, by using this interaction term, we can test the assumption that the effect of urbanization economies on local productivity in Italy varies moving from the South to the North, or from North-Western to North-Eastern regions.

Finally, Eq. (7) can be augmented by relaxing the *iid* assumption for the error term, that is assuming an error vector $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\varepsilon}}^2 \boldsymbol{\Lambda})$ with a covariance matrix $\boldsymbol{\Lambda}$ reflecting spatial error correlation as, for example, in Pinheiro and Bates (2000).

[5] Although this model is widely used in environmental studies and in epidemiology (see, i.a., Augustin et al. 2009), it is rarely considered for modeling economic data.

3.2 Parameter Estimation and Inference

Let us now discuss the issues concerning the estimation of model parameters in Eq. (7) and the related inference starting from the assumptions of an independent error structure and strict exogeneity of all explanatory variables. Omitting the subscript i, each *kth* univariate smooth term in Eq. (7) can be approximated by a linear combination of q_k known basis functions $b_{q_k}(x_k)$:

$$f_k(x_k) = \sum_{q_k} \beta_{q_k} b_{q_k}(x_k)$$

where β_{q_k} are unknown parameters to be estimated. To reduce mis-specification bias, $q_k's$ must be made fairly large. But this may generate a danger of over-fitting. As we shall clarify further on, by penalizing 'wiggly' functions when fitting the model, the smoothness of the functions can be controlled. Thus, a measure of 'wiggliness' $J \equiv \boldsymbol{\beta}' \mathbf{S} \boldsymbol{\beta}$ is associated with each k smooth function, with $\mathbf{S}$ a positive semidefinite matrix. Typically, the quadratic penalty term is equivalent to an integral of squared second derivatives of the function, for example $\int f''(x)^2 dx$, but there are other possibilities such as the discrete penalties suggested by Eilers and Marx (1996).

The penalized spline base-learners can be extended to two or more dimensions to handle interactions by using thin-plate regression splines or tensor products (Wood 2006a, Section 4.1.5). In the case of a tensor product, smooth bases are built up from products of 'marginal' bases functions. For example,

$$f_3(x_3, x_4) = \sum_{q_3} \sum_{q_4} \beta_{q_3,q_4} b_{q_3}(x_3) b_{q_4}(x_4)$$

A similar representation can be given for the smooth spatial trend surface, $h(no, e)$. Corresponding wiggliness measures are derived from marginal penalties (Wood 2006a). Moreover, it is worth mentioning that, when $f(x_3, x_4)$—or $h(no, e)$—is represented using a tensor product, the basis for $f(x_3) + f(x_4)$ is strictly nested within the basis for $f(x_3, x_4)$. Thus, in order to test for smooth interaction effects, we do not need to include in the model the two further terms $f(x_3)$ and $f(x_4)$.

In the case of a varying coefficient term like $f_4(x_1)\, l$, the basis functions $b_{q_4}(x_1)$ are premultiplied by a diagonal matrix containing the values of the interaction variable (l). Similarly, in the case of a spatially varying coefficient term like $h(no, e)\, x_1$, the basis functions $b_{q_{no}}(no) b_{q_e}(e)$ are premultiplied by a diagonal matrix containing the values of the interaction variable x_1.

Given the bases for each smooth term, Eq. (7) can be rewritten in matrix form as a large linear model,

$$\begin{aligned} \mathbf{y} &= \mathbf{X}^* \boldsymbol{\beta}^* + \Sigma_{q_1} \beta_{1q_1} b_{1q_1}(x_1) + \Sigma_{q_2} \beta_{2q_2} b_{2q_2}(x_2) + \ldots + \boldsymbol{\varepsilon} \\ &= \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \end{aligned} \tag{8}$$

where matrix $\mathbf{X}$ includes $\mathbf{X}^*$ and all the basis functions evaluated at the covariate values, while $\boldsymbol{\beta}$ contains $\boldsymbol{\beta}^*$ and all the smooth coefficient vectors, β_q.

As mentioned previously, the number of parameters for each smooth term in a semiparametric model must be large enough to reduce misspecification bias, but not too large to escape over-fitting. To solve this trade-off, we need to penalize lack of smoothness. Thus, starting from the assumption of exogeneity of all the *r.h.s.* variables, model (8) can be estimated by solving the following optimization problem

$$\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_k \lambda_k \boldsymbol{\beta}' \mathbf{S}_k \boldsymbol{\beta} \qquad w.r.t. \qquad \boldsymbol{\beta} \tag{9}$$

subject to any constraint associated with the bases plus any constraint needed to ensure that the model is identifiable. $\|.\|^2$ is the Euclidean norm and $\lambda_k \geq 0$ are the smoothing parameters that control the fit vs. smoothness trade-off. Employing a large number of basis functions yields a flexible representation of the nonparametric effect $f_k(.)$ where the actual degree of smoothness can be adaptively chosen by varying λ_k.[6]

Given smoothing parameters, λ_k, the solution to (9) is the following penalized least square estimator:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X} + \sum_k \lambda_k \mathbf{S}_k \right)^{-1} \mathbf{X}'\mathbf{y}$$

The covariance matrix of $\widehat{\boldsymbol{\beta}}$ can be derived from that of $\mathbf{y}$

$$V_{\widehat{\boldsymbol{\beta}}} = \sigma_{\boldsymbol{\varepsilon}}^2 \left(\mathbf{X}'\mathbf{X} + \sum_k \lambda_k \mathbf{S}_k \right)^{-1} \mathbf{X}'\mathbf{X} \left(\mathbf{X}'\mathbf{X} + \sum_k \lambda_k \mathbf{S}_k \right)^{-1}$$

If we also assume normality, that is $\boldsymbol{\varepsilon} \sim \mathcal{N}\left(0, \mathbf{I}_n \sigma_{\boldsymbol{\varepsilon}}^2\right)$, then

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left(E\left(\widehat{\boldsymbol{\beta}}\right), V_{\widehat{\boldsymbol{\beta}}} \right)$$

It has been observed, however, that frequentist confidence intervals based on the naive use of $\widehat{\boldsymbol{\beta}}$ and the corresponding covariance matrix perform quite poorly in terms of realized coverage probability (Wood 2006b). Thus, in practice, in additive models based on penalized regression splines, frequentist inference yields us to reject the null hypothesis too often. To overcome this problem, and following Wahba (1983) and Silverman (1985), Wood (2006a,b) has implemented a

[6]It is worth noticing that in expression (9), for interactive terms, the penalty matrix $\mathbf{S}_k$ usually depends on both interacting variables, and the associated λ_k will have two components allowing for different degrees of smoothing.

Bayesian approach to coefficient uncertainty estimation. This strategy recognizes that, by imposing a particular penalty, we are effectively including some prior beliefs about the likely characteristics of the correct model. This can be translated into a Bayesian framework by specifying a prior distribution for the parameters β. Specifically, Wood (2006b) shows that using a Bayesian approach to uncertainty estimation results in a Bayesian posterior distribution of the parameters

$$\boldsymbol{\beta}|\mathbf{y} \sim \mathcal{N}\left(E\left(\widehat{\boldsymbol{\beta}}\right), \sigma_{\varepsilon}^2\left(\mathbf{X}'\mathbf{X} + \sum_k \lambda_k \mathbf{S}_k\right)^{-1}\right)$$

This latter result can be used directly to calculate credible intervals for any parameter. Moreover, the credibility intervals derived via Bayesian theory are well behaved also from a frequentist point of view, i.e., their average coverage probability is very close to the nominal level $1-\alpha$, where α is the significance level.

3.3 Smoothing Parameter Selection: GCV Score Minimization

A crucial issue in the use of penalized regression splines within an additive semiparametric model is the selection of the smoothing parameters, λ_k, controlling the trade-off between fidelity to the data and smoothness of the fitted spline. How should these values be selected? There are two main approaches to identify the optimum smoothing parameters. First, we can use prediction error criteria, such as generalized cross validation (GCV), Akaike information criterion (AIC), Bayesian information criterion (BIC) and so on. Alternatively we can rewrite the penalized additive model as a mixed model by decomposing each smooth term into fixed effect and random effect components and estimate the model by ML or restricted maximum likelihood (REML), treating λ_k as variance parameters (see Sect. 4).

As for the first method, we may select the values of $\hat{\lambda}_k$ that minimize the GCV score:

$$GCV(\lambda_k) = \frac{n \parallel \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \parallel^2}{[n - tr(\mathbf{H})]^2}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X} + \sum \lambda_k \mathbf{S}_k)^{-1}\mathbf{X}'$ is the hat matrix for the model being fitted and its trace, $tr(\mathbf{H})$, gives the effective degrees of freedom *edf* (i.e., the number of identifiable parameters in the model). The *edf* is a general measure for the complexity of a function's estimate, which allows us to compare the smoothness, even for different types of effects (e.g. nonparametric versus parametric effects). If λ_k=0, then *edf* is equal to the size of the $\boldsymbol{\beta}$ vector minus the number of constraints (i.e., $edf = K$). Positive values of λ_k lead to an effective reduction of the number of parameters (i.e., $edf < K$). If λ_k is high, we have very few *edf*.

Actually, multiple smoothing parameter selection based on the minimization of the GCV score (10) is often computationally too demanding. To overcome this problem, Wood (2000) extended the 'performance iteration' method proposed by Gu and Wahba (1991) for automatically selecting multiple smoothing parameters to the case of computationally efficient low rank additive models based on penalized regression splines. First, the multiple smoothing parameter model fitting problem is re-written with an extra *overall* smoothing parameter (δ) controlling the trade-off between model fit and overall smoothness, while retaining smoothing parameters multiplying each individual penalty, which now control only the relative weights given to the different penalties. The following steps are then iterated: (1) given the current estimates of the relative smoothing parameters (λ_k/δ), estimate the overall smoothing parameter; and (2) given the overall smoothing parameter, update $log(\lambda_k)$ by Newton's method. In this way, the smoothing parameters for each smooth term in the model are chosen simultaneously and automatically as part of the model fitting. A drawback of this method is that it does not allow users to fix some smoothing parameters and estimate others or to bound smoothing parameters from below. Moreover, the method is not optimally stable numerically.

More recently, Wood (2004) proposed an improved (optimally stable) version of the the 'performance iteration' method which is more robust to collinearity or concurvity problems and which can deal with fixed penalties. The second issue is very important when fully automatic smoothing parameter selection results in one or more model terms clearly over-fitting and thus it is necessary to fix or bound smoothing parameters. The issue is also particularly relevant in geoadditive models, since the smooth function of spatial location ($h(no_i, e_i)$), which enter the model as a nuisance term—i.e. only to explain variability that cannot be explained by the covariates that are really of interest—is often estimated with bounded smoothing parameters, while the 'interesting' term are left with free smoothing parameters: in this way the 'interesting' covariates can be forced to do as much of the explanatory work as possible.

3.4 Semiparametric Spatial Autoregressive Geoadditive Models

Matrix $\mathbf{X}$ in model (8) may include spatial lags of the covariates, thus providing a Semiparametric Geoadditive Spatial Lag of $\mathbf{X}$ (SLX) Model (or SLX-Geo-AM). In other words, if $\widetilde{\mathbf{X}}$ is the matrix of regional characteristics, $\mathbf{X}$ includes both $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{W}_n\mathbf{X}}$, where $\mathbf{W}_n$ is a row-standardized spatial weights matrix. It is important to remark again, however, that the $\widetilde{\mathbf{W}_n X}$ only captures *local spatial externalities*. By replacing $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ with $(\mathbf{I}_n - \varrho\mathbf{W}_n)^{-1}\mathbf{X}$ and $(\mathbf{I}_n - \varrho\mathbf{W}_n)^{-1}\boldsymbol{\varepsilon}$, respectively, it is possible to model *global spillovers*.

The introduction of the spatial multiplier effect in the model yields a reduced form as $\mathbf{y} = (\mathbf{I}_n - \varrho \mathbf{W}_n)^{-1}\mathbf{X} + (\mathbf{I}_n - \varrho \mathbf{W}_n)^{-1}\boldsymbol{\varepsilon}$ and the structural form becomes a Semiparametric Spatial Autoregressive Geoadditive Model (SAR-Geo-AM):

$$y_i = \mathbf{X}_i^{*'}\boldsymbol{\beta}^* + \varrho \sum\nolimits_{j=1}^{n} w_{ij} y_j + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + f_4(x_{1i})\, l_i + \ldots + h(no_i, e_i) + \varepsilon_i \tag{10}$$

$$\varepsilon_i \sim iid\mathcal{N}(0, \sigma_{\varepsilon}^2) \qquad i = 1, \ldots, n$$

where, again, w_{ij} are the elements of $\mathbf{W}_n$, $\sum_{j=1}^{n} w_{ij} y_j$, which captures the spatial lag of the dependent variable (which always enters the model linearly), and ϱ is the spatial spillover parameter. This model was first proposed by Gress (2004) and Basile and Gress (2005) and then reformulated by Basile (2008, 2009), Basile et al. (2012), Montero et al. (2012), Mínguez et al. (2012), Su and Jin (2010) and Su (2012). It reflects the notion of a spatial correlation comprised of two parts: (a) a spatial trend due to unobserved regional characteristics, which is modeled by the smooth function of the coordinates, and (b) local and/or global spatial spillover effects, which can be modeled by including spatial lag terms of the independent and dependent variables. Su (2012) extends this model to allow for both heteroskedasticity and spatial dependence in the error term.

As mentioned in Sect. 2, the spatial lag term $\sum_{j=1}^{n} w_{ij} y_i$ and the error term ε_i are correlated. In order to deal with this endogeneity problem, the "control function" approach (Blundell and Powell 2003) can be used (Basile 2009). This is a simple two-step procedure. In the first step, an auxiliary semiparametric regression

$$\sum\nolimits_{j=1}^{n} w_{ij} y_i = \mathbf{X}_i^{*'}\boldsymbol{\beta}^* + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + f_4(x_{1i})\, l_i + \ldots + h(no_i, e_i) + \sum\nolimits_{m} g_m(Q_{mi}) + v_i \tag{11}$$

is considered, with $\mathbf{Q}_i$ a set of m conformable instruments,[7] and v_i a sequence of random variables satisfying conditional mean restrictions $E(v_i|\mathbf{Q}_i) = 0$.

The second step consists of estimating an additive model of the form:[8]

$$y_i = \mathbf{X}_i^{*'}\boldsymbol{\beta}^* + \varrho \sum\nolimits_{j=1}^{n} w_{ij} y_j + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + f_4(x_{1i})\, l_i + \ldots + h(no_i, e_i) + c(\hat{v}_i) + \varepsilon_i \tag{12}$$

[7] For example, in line with Kelejian and Prucha (1997), $\mathbf{Q}_i$ may contain an intercept, all exogenous terms included in the model and several orders of their spatial lags.

[8] Both first and second step equations can be estimated by using, for example, penalized least squares estimators.

Obviously, the endogeneity of any other continuously distributed regressor in model (10) can also addressed via the control function approach if valid instruments are available.[9] Finally, it is important to note that endogeneity problems arising from omitted variables (i.e., missing permanent characteristics that drive both the response variable and the covariates) can be ruled out since in model (10) we directly control for the effect of "first nature" characteristics by including the smooth interaction between latitude and longitude of the regional units of analysis.

In Sect. 4 we present a different possible solution to this simultaneity problem based on the REML estimation approach.[10] However, before leaving this discussion, an important issue remains to be introduced. Specifically, as in the parametric SAR model, in the semiparametric SAR also the estimated coefficients of parametric terms, as well as the estimated smooth effects of nonparametric terms, cannot be interpreted as marginal impacts of the explanatory variables on the dependent variable, due to the presence of a significant spatial autoregressive parameter (ϱ). Taking advantage of the results obtained for parametric SAR models (see Sect. 2), we can define similar algorithms for the semiparametric SAR model. Specifically, we can compute the total effect of variable x_k as

$$\hat{f}_k^{te_k}(x_k) = \Sigma_q \left[\mathbf{I}_n - \hat{\varrho}\mathbf{W}_n\right]_{ij}^{-1} b_{kq}(x_k)\hat{\beta}_{kq} \tag{13}$$

Finally, we can compute direct and indirect (or spillover) effects of smooth terms in semiparametric SAR as follows:

$$\hat{f}_k^{de_k}(x_k) = \Sigma_q \left[\mathbf{I}_n - \hat{\varrho}\mathbf{W}_n\right]_{ii}^{-1} b_{kq}(x_k)\hat{\beta}_{kq} \tag{14}$$

$$\hat{f}_k^{ie_k}(x_k) = \hat{f}_k^{te_k}(x_k) - \hat{f}_k^{de_k}(x_k) \tag{15}$$

4 Semiparametric Geoadditive Models as Mixed Models

Semiparametric models presented in the previous section can also be expressed as mixed models. Consequently, it is possible to estimate all the parameters of these models using restricted maximum likelihood methods (REML). In Sect. 4.1 we present some ways to deal with general semiparametric models using mixed models.

[9] The requirement that the endogenous regressor be continuously distributed is the most important limitation of the applicability of the control function approach to estimation of nonparametric and semiparametric models with endogenous regressors.

[10] It is important to mention that a semiparametric spatial lag model has also been proposed within a partial linear framework. For example, Su and Jin (2010) develop a profile quasi-maximum likelihood estimator for the partially linear spatial autoregressive model which combines the spatial autoregressive model and the nonparametric (local polynomial) regression model. Furthermore, Su (2012) proposes a semiparametric GMM estimator of the SAR model under weak moment conditions which allows for both heteroskedasticity and spatial dependence in the error terms.

Moreover, it is possible to estimate the whole set of parameters (including those for smoothing and interaction between variables) using REML. Sect. 4.2 shows how this methodology can be applied to estimate the parameters of *SAR-Geo-AM* model in a single step.

4.1 Model Specification and REML Estimation

The estimation of model (8) can be based on the reparameterization of such a model in the form of a mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U} + \boldsymbol{\varepsilon} \quad \mathbf{U} \sim \text{ i.i.d. } N(\mathbf{0}, \mathbf{G}) \quad \boldsymbol{\varepsilon} \sim \text{ i.i.d. } N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}) \tag{16}$$

where $\mathbf{G}$ is a block-diagonal matrix, which depends on both $\sigma_{u_k}^2$ and σ_ε^2 variances. The smoothing parameters are defined by the ratios $\lambda_k = \frac{\sigma_\varepsilon^2}{\sigma_{u_k}^2}$. Again, matrix $\mathbf{X}$ may include parametric components such as the intercept, continuous covariates and categorical covariates.

This reparameterization consists in postmultiplying $\mathbf{X}$ and premultiplying $\boldsymbol{\beta}$ in model (8) by an orthogonal matrix resulting from the singular value decomposition of the penalty matrices $\mathbf{S}_k$ (Wand 2003; Lee and Durbán 2011; Wood et al. 2012). Therefore, the type of penalizations determines the transformation matrix and, thus, the fixed and random effects obtained in the mixed model. The resulting coefficients associated with the fixed effects are not penalized, while those associated with the random effects are penalized. The penalization of random effects is given by the variance–covariance matrix of these coefficients.

It is worth pointing out that when the model is a pure additive model $\mathbf{y} = \sum_{k=1}^{K} f(x_k) + \boldsymbol{\varepsilon}$ (i.e. there are no interaction terms), $\mathbf{G}$ is block-diagonal, each block matrix $\mathbf{G}_k$ depending only on λ_k, the smoothing coefficient associated to each variable x_k. Thus, model (16) becomes a variance components model that can be estimated by using standard software on the topic. When the model contains interaction terms, it is no longer a pure additive model. Therefore, each block $\mathbf{G}_k$ depends on more than one smoothing coefficient λ_k, except in the isotropic case,[11] where coefficients λ_k are the same for all variables (Wood et al. 2012; Lee and Durbán 2011). As a consequence, the resulting mixed model is not an orthogonal variance component model and standard software cannot be used to estimate it.

A recent reparameterization, proposed by Wood et al. (2012), allows us to express a semiparametric model including additive and interaction effects as a mixed model

[11] For the sake of clarity, isotropy means that the degree of smoothness is the same for all the covariates, that is, the degree of flexibility in all of them is the same. Nevertheless, the usual situation in real cases is anisotropy, since the covariates are usually measured in different units of measure or, in the case of equal measurement units (e.g. spatial location variables), the variability of such covariates differing greatly.

with orthogonal variance components, allowing for different degrees of smoothing for the variables that interact using only one smoothing coefficient for each term. An alternative reparameterization from a P-Spline approach with a B-Spline basis and penalization matrices for the basis coefficients based on discrete differences is considered in Eilers and Marx (1996) and Lee and Durbán (2011). Two other interesting reparameterizations are based on (a) a truncated polynomial basis and ridge penalizations (Ruppert et al. 2003), and (b) on a thin plate regression splines basis and penalizations based on the integral of the second derivatives of the spline functions (Wood 2003). These last three alternatives cannot be estimated with standard software on mixed models when the interactions between the variables are considered (except in the isotropic case).

Once the mixed model is defined, the parameters associated to fixed ($\boldsymbol{\beta}$) and random effects (λ_k and σ_ε^2) can be estimated by using a ML algorithm. If the noise term follows a Gaussian distribution, the log-likelihood function is given by:

$$\log L(\boldsymbol{\beta}, \lambda_1, \cdots, \lambda_K, \sigma_\varepsilon^2) = \text{constant} - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

where $\mathbf{V} = \mathbf{ZGZ}' + \sigma_\varepsilon^2\mathbf{I}$ and the smoothing parameters λ_k are included in $\mathbf{V}$.

However, the ML estimates are biased since this method does not take into account the reduction in the degrees of freedom due to the estimation of the fixed effects. The restricted maximum likelihood (REML) method can be used to solve the problem. The REML method looks for the linear combinations of the dependent variable that eliminates the fixed effects in the model (McCulloch et al. 2008). In this case the objective function to maximize is given by:

$$\begin{aligned}\log L_{\mathrm{R}}(\lambda_1, \cdots, \lambda_K, \sigma_\varepsilon^2) = \text{constant} &- \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| \\ &- \frac{1}{2}\mathbf{y}'\left(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\right)\mathbf{y}\end{aligned}$$

An estimation of the variance components parameters can be obtained after maximizing $\log L_{\mathrm{R}}(.)$. In a second step, the estimates of $\boldsymbol{\beta}$ and $\mathbf{U}$ are given by (McCulloch et al. 2008):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$$

$$\hat{\mathbf{U}} = \hat{\mathbf{G}}\mathbf{X}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Finally, the estimated values of the observed variable can be obtained as:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{U}}$$

To build confidence intervals for the estimated values, an approximation of the variance–covariance matrix of the estimation error is given by $V(\mathbf{y}-\hat{\mathbf{y}}) = \sigma_\varepsilon^2\mathbf{H}$ where, as shown previously in the GCV method, $\mathbf{H}$ is the hat matrix of the model (Ruppert et al. 2003). For the mixed model, it can be proved that:

$$\mathbf{H} = \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{pmatrix}$$

Recently Wood (2011) has proposed a Laplace approximation to obtain an approximated REML or ML for any generalized linear model, which is suitable for efficient direct optimization. Simulation results indicate that these novel REML and ML procedures offer, in most cases, significant gains (in terms of mean-square error) with respect to GCV or AIC methods.

4.2 Semiparametric Spatial Autoregressive Geoadditive Models as Mixed Models

In a mixed-model form the SAR-Geo-AM can be expressed as:

$$\mathbf{y} = \varrho\mathbf{W}_n\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U} + \boldsymbol{\varepsilon} \quad \mathbf{U} \sim \text{ i.i.d. } N(\mathbf{0}, \mathbf{G}) \quad \boldsymbol{\varepsilon} \sim \text{ i.i.d. } N(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$$

In reduced form we have:

$$\mathbf{y} = \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \mathbf{A}\mathbf{Z}\mathbf{U} + \mathbf{A}\boldsymbol{\varepsilon} \tag{17}$$

where $\mathbf{A} = (\mathbf{I} - \varrho\mathbf{W}_n)^{-1}$.

As pointed out in Montero et al. (2012) and Mínguez et al. (2012), the log-REML function for model (17) is:

$$\log L_{\mathrm{R}}(\rho, \lambda_1, \cdots, \lambda_K, \sigma_\varepsilon^2) = \text{constant} - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \log|\mathbf{A}|$$
$$-\frac{1}{2}\mathbf{y}'\mathbf{A}'\left(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\right)\mathbf{A}\mathbf{y}$$

As usual, $\log L_{\mathrm{R}}(.)$ is maximized with respect to the parameter vector $(\lambda_1, \cdots, \lambda_K, \sigma_\varepsilon^2)'$. Note that the maximization process requires the computation of the log-determinant of matrix $\mathbf{A}$, a dense $n \times n$ inverse matrix depending on ϱ. As a consequence, the maximization of such a function constitutes a challenging task. Nevertheless, to evaluate $\mathbf{A}$ for different values of ϱ when n is large, it is possible to use Monte Carlo procedures (LeSage and Pace 2009).

Finally, fixed and random effects can be estimated as:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})\mathbf{X}'\hat{\mathbf{V}}^{-1}\hat{\mathbf{A}}\mathbf{y}$$
$$\hat{\mathbf{U}} = \hat{\mathbf{G}}\mathbf{X}'\hat{\mathbf{V}}^{-1}(\hat{\mathbf{A}}\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Unlike model (16), the spatial lag model (17) cannot be estimated by using standard software regardless of the type of reparameterization used to express it as a mixed model.[12]

5 Empirical Applications

The scientific contributions to the empirical literature on regional economic dynamics and economic geography may be roughly classified into four main categories depending on whether they apply (a) parametric non-spatial models, (b) parametric spatial models, (c) semiparametric non-spatial models and, (d) semiparametric spatial geoadditive models. In this section we mention examples for each category, focusing on empirical studies on regional and urban growth and on economic geography. In particular, in Sect. 5.2 we present an application of the SAR-Geo-AM to regional growth data.

5.1 Studies Based on Parametric Models

The main difference between the first two categories lies in the treatment of spatial effects. Studies belonging to the first category are based on the assumption of independence and, thus, spillover or feedback effects are ruled out. For example, Magrini (2004) provides a review of empirical studies on regional growth and convergence, which mainly focus on testing the β-convergence hypothesis, by regressing the so-called linear Barro's equation, derived from the neoclassical Solow model and based on the assumption of cross-region independence. Similarly, in the field of urban economics, a large number of studies have investigated the relationship between agglomeration externalities and local economic growth, ignoring regional interdependence (Rosenthal and Strange 2004; Beaudry and Schiffauerova 2009). Moreover, several studies have put into empirical test selected propositions of New Economic Geography (NEG) models, such as the prediction of a positive relationship between local wages and market potential (home market effect). However, none of the empirical studies on the NEG wage equation take

[12]Nevertheless, there are some *R* codes using *spdep* package available from Montero et al. (2012) and Mínguez et al. (2012).

spatial dependence into account (Head and Mayer 2004; Brakman et al. 2006; Redding 2010).

Recent studies (López-Bazo et al. 2004; Pfaffermayer 2009; Ertur and Koch 2007, 2011) have shown that spatial technological interdependence can be explicitly modeled in multi-region exogenous and endogenous growth frameworks to account for neighborhood effects in growth and convergence processes. These studies have provided sound theoretical foundations for the specific form taken by spatial autocorrelation in econometric growth models. Thus, they have stimulated the empirical assessment of the existence of neighboring effects in regional growth (Rey and Gallo 2009).

In the same spirit, Behrens and Thisse (2007) stressed the importance of extending the two-region NEG models into multi-region models:

> when there are just two regions, there is only one way in which these regions can interact, namely directly; whereas with three regions, there are two ways in which these regions can interact, namely directly and indirectly. In other words, in multiregional systems the so-called 'three-ness effect' enters the picture and introduces complex feedbacks into the models, which significantly complicates the analysis. Dealing with these spatial interdependencies constitutes one of the main theoretical and empirical challenges NEG and regional economics will surely have to face in the future (p. 461).

Multi-region NEG models have been empirically tested by Bode and Mutl (2010) and Fingleton (2006). In particular, Bode and Mutl (2010) derive a reduced-form linearized wage equation, which is a simple SAR model of order one in short-run deviations of local wages from their equilibrium values. This model relates these deviations in each region to the weighted sum of the deviations in all regions. The spatial weights are bilateral elasticities of the wage rate in one region with respect to the wage rate in another region. Estimation of this model thus becomes a test of whether or not local wage shocks propagate through the system of observed regional wages in the way predicted by the NEG model.

While the insights of these empirical studies employing sophisticated spatial econometric techniques are valuable to measure spatial spillover and feedback effects, their foundation on a priori functional form specification limits the scope of these methods in uncovering the process dictating regional dynamics.

5.2 Studies Based on Semiparametric Models

In this section, we mention examples of recent empirical works that adopt semiparametric additive models in regional science and economic geography. First, it is important to acknowledge the existence of several studies using semiparametric methods to identify nonlinearities, and of a few studies using spatial lag semiparametric models to tackle both spatial dependence and nonlinearities.

Following the cross-country growth literature (Durlauf et al. 2005), an emerging issue in regional growth analyses has been the evidence of strong nonlinearities in regional growth models (Fotopoulos 2012; Azomahou et al. 2011). Using

semiparametric methods, these studies have uncovered the existence of significant nonlinearities across an array of variables within cross-region growth regressions. Moreover, in the field of urban growth, Basile et al. (2013) propose a semiparametric geoadditive model to identify important nonlinearities in the relationship between local industry structure (e.g. population density). Although these studies are able to relax functional form assumptions, their consistency still depends on restrictive assumptions about interregional independence. Finally, the need to consider jointly spatial dependence and nonlinearities has been raised by Gress (2004), Basile and Gress (2005), Basile (2008, 2009) and Basile et al. (2012).

As an example of the application of a semiparametric spatial lag geoadditive model (SAR-Geo-AM), we report the results of the estimation of a regional growth regression model on a sample of 249 NUTS2 regions belonging to the enlarged Europe (EU27). The dependent variable, y, is the per-worker income growth rate ($y = (ln(p_T) - ln(p_0))/T$), computed for the 1990–2004 period. The covariates are the rates of investment in physical and human capital ($ln(s_k)$ and $ln(s_h)$, respectively), initial conditions ($ln(p_0)$) and the effective depreciation rate ($ln(n + g + \delta)$), with n the working-age population growth rate, g the common exogenous technology growth rate and δ the rate of depreciation of physical capital assumed identical in all economies. Basic data to measure these variables come from the EUROSTAT Regio and Cambridge Econometrics databases, which include information on real gross value added, employment, investment and tertiary education (for further details, see Basile et al. 2012). The estimated model is

$$\begin{aligned} y_i = \beta_0 + \varrho \sum\nolimits_{j=1}^{n} w_{ij} y_j + f_1\left(ln(p_0)_i\right)) + f_2\left(ln(s_k)_i\right)) \\ + f_3\left(ln(s_h)_i\right)) + f_4\left(ln(n + g + \delta)_i\right)) + h\left(no_i, e_i\right) + \varepsilon_i \end{aligned} \tag{18}$$

$$\varepsilon_i \sim iid\mathcal{N}(0, \sigma_{\varepsilon}^2) \qquad i = 1, \ldots, n$$

The matrix $\mathbf{W}_n$ used to estimate this model has been selected among a number of inverse-distance spatial weights matrices. The model has been estimated using spline–based penalized regression smoothers which allow for automatic and integrated smoothing parameters selection via *GCV*. A control function approach has been used in order to deal the endogeneity of the spatial lag term, and the spatial lags of the exogenous variables have been used as valid instruments.

The F tests for the overall significance of the smooth terms have p values lower than 0.05 in all cases, while the number of *edf* suggests that the relationships between regional growth and its determinants are far from being linear. Figure 1a–d show the fitted univariate smooth functions, alongside Bayesian credibility intervals at the 95 % level of significance. The value of the spatial autocorrelation parameter ϱ is equal to 0.88 and statistically significant at 1 %, confirming the role of spatial frictions in the interregional diffusion of technological spillovers.

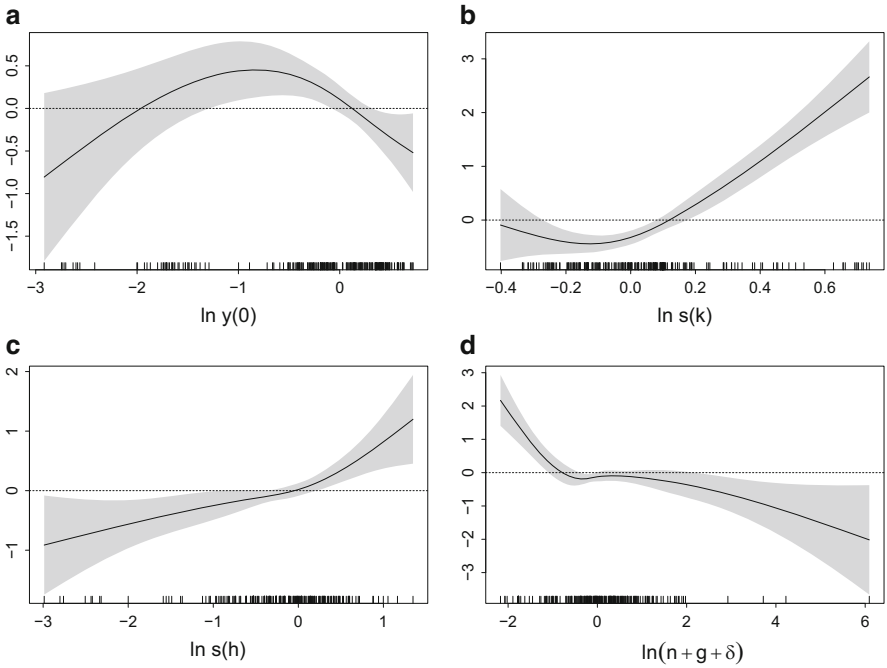


Fig. 1 Smooth effects of growth determinants. (**a**) Initial conditions. (**b**) Physical capital accumulation rate. (**c**) Human capital accumulation rate. (**d**) Effective rate of depreciation. *Solid lines* represent smooth functions of each term, alongside Bayesian confidence intervals (*shaded grey areas*) at the 95 % level of significance. In each plot, the vertical axis displays the scale of the estimated smooth function, while the horizontal ones report the scale of each determinant (in deviations to the EU average). Rug plot along the horizontal axis represents observed data

A hump-shaped relationship between growth and initial conditions emerges. Specifically, a diverging behavior characterizes the group of Eastern regions (45 regions), while Western regions maintain a conditional predicted convergence path. The assumption of identical speed of convergence is therefore rejected. Nonlinearities in the effects of gross physical investments are also clearly detected. Specifically, an increase in the saving rate is associated with an increase in growth rates only when the saving rate is above the EU average. A similar threshold effect is also evident in the smooth effect of human capital investments. The influence of the employment growth rate on regional growth is negative, although the effect is not homogeneous across the sample. Moreover, Fig. 2 displays the effect of the smooth interaction between latitude and longitude. It can be observed that, ceteris paribus, some North-Eastern and some North-Western regions (mainly the UK and Ireland) have higher predicted growth rates.

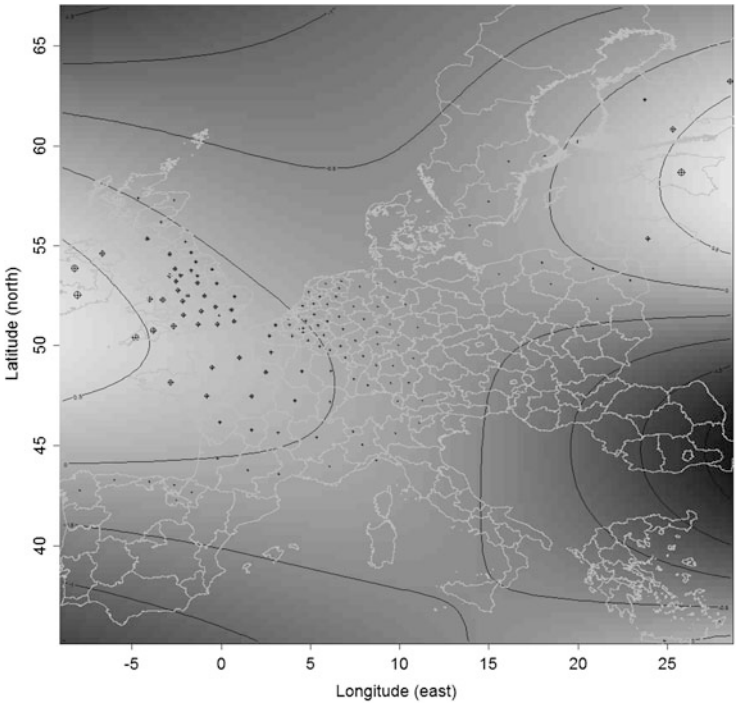


Fig. 2 Spatial trend surface

Finally, it is important to remark again that the estimated smooth effects of nonparametric terms cannot be interpreted as marginal impacts of the explanatory variables on the dependent variable. Therefore, using Eqs. (13), (14) and (15), we have computed direct, indirect and total smooth effects. Actually, these effects are not smooth at all over the domain of variable x_k due to the presence of the spatial multiplier matrix in these algorithms. A wiggly profile of direct, indirect and total effects would appear even if the model were linear.[13] Therefore, in the spirit of this paper, we have applied a spline smoother to obtain smooth curves (see Fig. 3). Briefly commenting these evidences, we first note that the shape of direct effects is very similar to the one displayed in Fig. 1, which means that the feedback effect is rather negligible. Moreover, indirect effects (spillover effects) are always lower that direct effects.

[13]This kind of heterogeneity is called interactive heterogeneity by Ertur and Koch (2011), and this is why scholars usually compute average marginal effects for parametric SAR and SDM models.

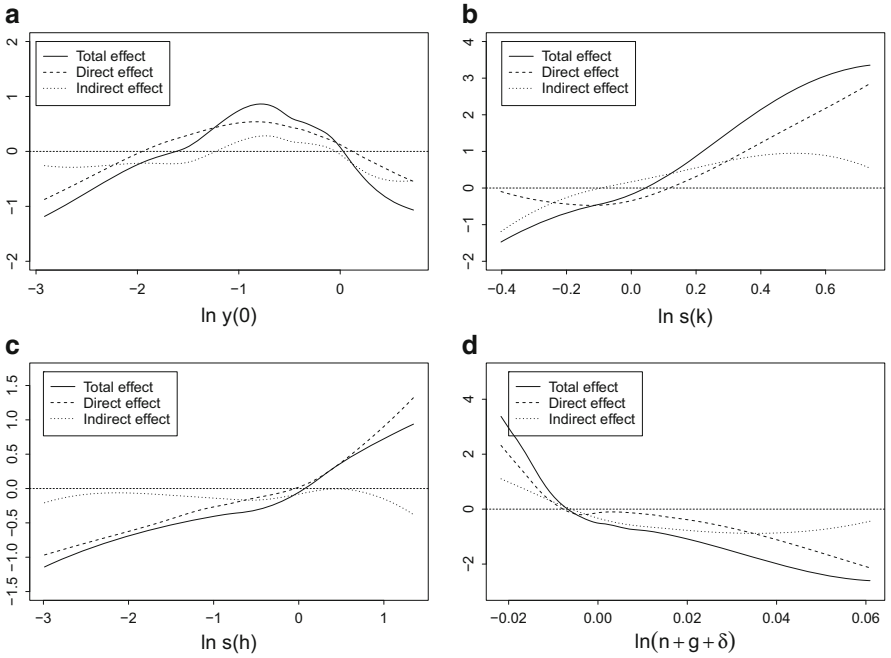


Fig. 3 Direct, indirect and total smooth effects [see Eqs. (13), (14) and (15)]. (**a**) Initial conditions. (**b**) Physical capital accumulation rate. (**c**) Human capital accumulation rate. (**d**) Effective rate of depreciation

6 Concluding Remarks

In this paper we have reviewed recently developed spatial lag semiparametric geoadditive models and presented some applications of them in the fields of regional science and economic geography. These methods play a prominent role in those context in which the theory suggests the existence of spatial interdependence and heterogeneous behavior of the spatial units. Natural directions in which these methods can be extended are a specification for longitudinal data and, eventually, a dynamic framework. Also testing spatial autocorrelation in the residuals in both non-spatial and spatial lag semiparametric geoadditive models is an important task.

Acknowledgements This paper has been presented at the EU-COST Meeting in Lisboa. We thank all participants for useful comments. We are also grateful to a referee that with his/her comments helped us to reformulate the analysis. We are responsible for any remaining errors. The work of Román Mínguez was supported by the research project MTM-2011-28285-C02-C2 from the Spanish Government's Ministry of Economy and Competitiveness. Jesús Mur likes to thank the financial support of the research project ECO2012-36032-C03-01 from the Spanish Government's Ministry of Economy and Competitiveness.

References

Anselin, L. (2003). Spatial externalities, spatial multipliers and spatial econometrics. *International Regional Science Review, 26*, 153–166.

Augustin, N., Musio, M., Wilpert, K. V., Kublin, E., Wood, S., & Schumacher, M. (2009). Modeling spatio-temporal forest health monitoring data. *Journal of the American Statistical Association, 104*, 899–911.

Azomahou, T., Ouardighi, J. E., Nguyen-Van, P., & Pham, T. (2011). Testing convergence of European regions: A semiparametric approach. *Economic Modelling, 28*, 1202–1210.

Basile, R. (2008). Regional economic growth in Europe: A semiparametric spatial dependence approach. *Papers in Regional Science, 87*, 527–544.

Basile, R. (2009). Productivity polarization across regions in Europe: The role of nonlinearities and spatial dependence. *International Regional Science Review, 32*, 92–115.

Basile, R., Capello, R., & Caragliu, A. (2012). Technological interdependence and regional growth in Europe. *Papers in Regional Science, 91*, 697–722.

Basile, R., Donati, C., & Pittiglio, R. (2013). *Industry structure and employment growth: Evidence from semiparametric geoadditive models*. Mimeo.

Basile, R., & Gress, B. (2005). Semi-parametric spatial auto-covariance models of regional growth behavior in Europe. *Region et Developement, 21*, 93–118.

Beaudry, C., & Schiffauerova, A. (2009). Who's right, Marshall or Jacobs? The localization versus urbanization debate. *Research Policy, 38*, 318–337.

Behrens, K., & Thisse, J. (2007). Regional economics: A new economic geography perspective. *Regional Science and Urban Economics, 37*, 457–465.

Blundell, R., & Powell, J. (2003). Endogeneity in nonparametric and semiparametric regression models. In *Advances in economics and econometrics theory and application*. Cambridge: Cambridge University Press.

Bode, E., & Mutl, J. (2010). *Testing nonlinear new economic geography models*. Reihe Ökonomie, Working Paper No. 253. http://hdl.handle.net/10419/45569.

Brakman, S., Garretsen, H., & Schramm, M. (2006). Putting new economic geography to the test: Free-ness of trade and agglomeration in the EU regions. *Regional Science and Urban Economics, 36*, 613–635.

Durlauf, S., Johnson, P., & Temple, J. (2005). Growth econometrics. In *Handbook of economic growth* (pp. 555–677). New York: North Holland.

Eilers, P., & Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science, 11*, 89–121.

Elhorst, P. (2010). Spatial panel data models. In *Handbook of applied spatial analysis* (pp. 377–407). Berlin: Springer.

Elhorst, P. (2012). Dynamic spatial panels: Models, methods, and inferences. *Journal of Geographical Systems, 14*, 5–28.

Ertur, C., & Gallo, J. L. (2009). Regional growth and convergence: Heterogenous reaction versus interaction spatial econometric approaches. In *Handbook of regional growth and development theories* (pp. 374–388). Cheltenham: Edward Elgar.

Ertur, C., & Koch, W. (2007). Growth, technological interdependence and spatial externalities: Theory and evidence. *Journal of Applied Econometrics, 22*, 1033–1062.

Ertur, C., & Koch, W. (2011). A contribution to the schumpeterian growth theory and empirics. *Journal of Economic Growth, 16*, 215–255.

Fingleton, B. (2004). Regional economic growth and convergence: Insights from a spatial econometric perspective, In *Advances in spatial econometrics* (pp. 397–432). Berlin: Springer.

Fingleton, B. (2006). The new economic geography versus urban economics: An evaluation using local wage rates in great Britain. *Oxford Economic Papers, 58*, 501–530.

Fingleton, B., & López-Bazo, E. (2006). Empirical growth models with spatial effects. *Papers in Regional Science, 85*, 177–198.

Fotheringham, A., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression*. Chichester: Wiley.

Fotopoulos, G. (2012). Nonlinearities in regional economic growth and convergence: The role of entrepreneurship in the European union regions. *The Annals of Regional Science, 48*, 719–741.

Gress, B. (2004). *Using semi-parametric spatial autocorrelation models to improve hedonic housing price prediction*. Mimeo. Department of Economics, University of California.

Gu, C., & Wahba, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM Journal on Scientific and Statistical Computing, 12(2)*, 383–398.

Hastie, T., & Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman and Hall.

Head, D., & Mayer, T. (2004). The empirics of agglomeration and trade. In *Handbook of urban and regional economics* (Vol. 4, pp. 2609–2669). New York: North-Holland.

Henderson, V. (1997). Externalities and industrial development. *Journal of Urban Economics, 42*, 449–470.

Holly, S., Pesaran, H., & Yamagata, T. (2010). A spatio-temporal model of house prices in the US. *Journal of Econometrics, 158*, 160–173.

Kapoor, M., Kelejian, H., & Prucha, I. (2007). Panel data models with spatially correlated error components. *Journal of Econometrics, 140*, 97–130.

Kelejian, H., & Prucha, I. (1997). Estimation of spatial regression models with autoregressive errors by two-stage least squares procedures: A serious problem. *International Regional Science Review, 20*, 103–111.

Kelejian, H., & Prucha, I. (2001). On the asymptotic distribution of Moran I test statistic with applications. *Journal of Econometrics, 104*, 291–257.

Kelejian, H., & Robinson, D. (2004). The influence of spatially correlated heteroskedasticity on tests of spatial correlation. In *Advances in spatial econometrics: Methodology, tools and applications* (pp. 79–97). Berlin: Springer.

Kneib, T., Hothorn, T., & Tutz, G. (2009). Variable selection and model choice in geoadditive regression models. *Biometrics, 65*(2), 626–634.

Krugman, P. (1993). First nature, second nature, and metropolitan location. *Journal of Regional Science, 33*, 129–144.

Lee, D., & Durbán, M. (2011). P-spline ANOVA type interaction models for spatio-temporal smoothing. *Statistical Modelling, 11*, 49–69.

Lee, L. (2004). Asymptotic distribution of quasi-maximum likelihood estimators for spatial econometric models. *Econometrica, 72*, 1899–1926.

LeSage, J., & Pace, K. (2009). *Introduction to spatial econometrics*. Boca Raton: CRC Press.

Liu, X., Lee, L., & Bollinger, C. (2007). An efficient GMM of spatial autoregressive models. *Journal of Econometrics, 159*, 303–319.

Lloyd, C. (2011). *Local models for spatial analysis* (2nd ed.). Boca Raton: CRC Press.

López-Bazo, E., Vayá, E., & Artís, M. (2004). Regional externalities and growth: Evidence from European regions. *Journal of Regional Science, 44*, 43–73.

Magrini, S. (2004). Regional (di)convergence. In *Handbook of regional and urban economics* (pp. 2741–2796). New York: North-Holland.

McCulloch, C., Searle, S., & Neuhaus, J. (2008). *Generalized, linear, and mixed models* (2nd ed.). Chichester: Wiley Series in Probability and Statistics.

McMillen, D. (2003). Spatial autocorrelation or model misspecification? *Interregional Regional Science Review, 26*, 208–217.

Mínguez, R., Durbán, M., Montero, J., & Lee, D. (2012). Competing spatial parametric and non-parametric specifications. Mimeo.

Montero, J., Mínguez, R., & Durbán, M. (2012). SAR models with nonparametric spatial trends. A P-spline approach. *Estadística Española, 54*, 89–111.

Mur, J., López, F., & Angulo, A. (2009). Testing the hypothesis of stability in spatial econometric models. *Papers in Regional Science, 88*, 409–444.

Pace, K., & LeSage, J. (2004). Spatial autoregressive local estimation. In *Spatial econometrics and spatial statistics* (pp. 31–51). Basingstoke: Palgrave Macmillan.

Paci, R., & Usai, S. (2008). Agglomeration economies, spatial dependence and local industry growth. *Revue D'Economie Industrielle, 123*, 1–23.

Partridge, M., Boarnet, M., Brakman, S., & Ottaviano, G. (2012). Introduction: Whither spatial econometrics. *Journal of Regional Science, 52*, 167–171.

Pfaffermayer, M. (2009). Conditional β and σ-convergence in space: A maximum likelihood approach. *Regional Science and Urban Economics, 39*, 63–78.

Pinheiro, J., & Bates, D. (2000). *Mixed-effects Models in S and S-PLUS*. New York: Springer.

Pinkse, J., Slade, M., & Brett, C. (2002). Spatial price competition. A semiparametric approach. *Econometrica, 70*, 1111–1153.

Redding, S. (2010). The empirics of new economic geography. *Journal of Regional Science, 50*, 297–311.

Rey, S., & Gallo, J. L. (2009). Spatial analysis of economic convergence. In *Handbook of econometrics. Applied econometrics* (Vol. II, pp. 1251–1293). Basingstoke: Palgrave Macmillan.

Rosenthal, S. S., & Strange, W. (2004). Evidence on the nature and sources of agglomeration economies. In *Handbook of urban and regional economics* (pp. 2119–2171). New York: North-Holland.

Ruppert, D., Wand, M., & Carroll, R. (2003). *Semiparametric regression*. Cambridge: Cambridge University Press.

Silverman, B. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B, 47*, 1–53.

Su, L. (2012). Semiparametric GMM estimation of spatial autoregressive models. *Journal of Econometrics, 167*, 543–560.

Su, L., & Jin, S. (2010). Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models. *Journal of Econometrics, 157*, 18–33.

Wahba, G. (1983). Bayesian confidence intervals for the cross validated smoothing spline. *Journal of the Royal Statistical Society. Series B, 45*, 133–150.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics, 18*, 223–249.

Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 65*, 95–114.

Wood, S. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association, 99*, 673–686.

Wood, S. (2006a). *Generalized additive models. An introduction with R*. London: Chapman and Hall.

Wood, S. (2006b). On confidence intervals for generalized additive models based on penalized regression splines. *Australian and New Zealand Journal of Statistics, 48*, 445–464.

Wood, S. (2011), Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 73*, 3–36.

Wood, S., Scheipl, F., & Faraway, J. (in press). Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing*. doi:10.1007/s11222-012-9314-z.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62*(2), 413–428.

Diffusion of Growth and Cycles in Continuous Time and Space

Tönu Puu

Abstract Spatial economics can be formatted in two different ways. One can see geographical space as a set of locations connected by arcs. Or one can see it as a continuous plane in two dimensions. Activities, such as production and consumption can be represented in both, and so can flows, of trade or migrants. When space is the continuous plane such flows are represented by continuous fields as in physics. The continuous outlook was the traditional both in economics and geography. However, it does not so easily lend itself to computations and fitting to actual data as in the discrete representation. The disadvantage of seeing space as a matrix of locations and connecting arcs is that space, in the sense of geometric shape and size slips out. In the twentieth century there were presented two ingenious economic models using continuous space, Hotelling (*A mathematical theory of migration*, MA Thesis, University of Washington, reprinted in *Environment and Planning A, 10*, 1223–1239) who dealt with migration, and Beckmann (1952) who dealt with trade and pricing in a spatially dispersed market. Here we take a simpler case dealing with the diffusion of growth and business cycles in continuous geographical space.

1 Introduction

The most cited business cycle model is, no doubt, due to Samuelson (1939). See also Hansen (1951). It is based on the multiplier and the principle of acceleration, and was later extended to a nonlinear format by Hicks (1950). Unlike earlier theories that used different explanations for upswing and downturn (see von Haberler 1937) it is self-contained, relating just national income and its components, consumption and investment. As it does not include anything monetary, it is basically Keyensian in appearance, even more Keynesian than Keynes's General Theory 1936 itself Keynes (1936); as there is no mention of the rate of interest, nor of the quantity of money. See Puu and Sushko (2006).

T. Puu (✉)
CERUM, Umeå University, 90187 Umeå, Sweden
e-mail: tonu.puu@cerum.umu.se

P. Commendatore et al. (eds.), *Complexity and Geographical Economics*,
Dynamic Modeling and Econometrics in Economics and Finance 19,
DOI 10.1007/978-3-319-12805-4_5

The model is cast in discrete time, which is natural with respect to its Keynesian background. National accounting emerged with Keynesian macroeconomics, and, for the first time in history, it became an operational concept in stead of a theoretical construct, such as interest on national wealth [the total of discounted expected future income streams on all assets in the economy (Lindahl 1939)].

Through delays and feedback the Samuelson model became a second order simple oscillator. Due to the linear structure there were only two possibilities for oscillations; either damped or explosive. None was good for producing sustained bounded amplitude oscillations. In the case of damping, some external disturbance had to be added (Frisch 1933), in the explosive case external bounds, such as the Hicksian "floor" and "ceiling" (Hicks 1950) had to be provided; though in his review Duesenberry showed that floor alone would do (Duesenberry 1950).

It may be noted that in modelling growth, Harrod (1948) set the tradition of using continuous time, as probably there was no urge to relate the model to anything observable. Ironically, Harrod emphasized that the balanced growth rate he defined, was unstable; and that any deviation of initial conditions from it would just set the system to some completely divergent trajectory. This was not understood by his followers.

1.1 Discrete Versus Continuous Time

To model what happens in such cases through a higher order process in continuous time, was left to Phillips (1954). See also Allen (1956). What time lags do in discrete time, adaptation can do in continuous time. The choices between discrete or continuous modelling have hence been due to historical coincidence rather than to necessity in terms of subject matter.

In what follows, we explore what can be modelled using continuous time in models of economic growth and business cycles. However, we include a spatial extension of the economy, conceived as continuous two-dimensional geographical space, where the locations are linked through a linear interregional trade multiplier, quite as was the tradition in Keynesian models of the open economy. See Metzler (1950). Consumption plus investment, plus export surplus then equal income.

Given a constant interregional import propensity, imports (like local consumption) then are proportional to local income, whereas exports are proportional to incomes in the surrounding space. Assuming the import propensity is constant also over space, export surplus would then equal this constant multiplied by the difference of income in the surrounding points and the point itself. We only need to define a reasonable measure of such spatial differences of conditions in a point and its surroundings in the continuous two-dimensional plane.

1.2 On Time and Space

Economics almost never dealt with such issues. Fortunately there exist concepts to borrow from mathematical physics for linear diffusion in space. For a start we have to realize that there is a fundamental difference between time and space, even when space is the one-dimensional line (which it too often remains throughout models of space economies even when it is said that one dimension is just a first step towards a full analysis of the geographical pla(i)ne).

Time implies a forward direction, and there is a difference between past and future. If a variable has positive time derivative, it increases as time passes and decreases when facing the past.

In space any location interacts with both left and right on equal terms (unless we are in the rare case of a boundary point where special conditions may apply). We can compare the change when leaving a point to the right through the first derivative, when leaving it to the left through its *negative*. Taking both directions in account, it is the *difference* of the right and left derivatives that counts, which in the limit it goes to the second derivative. Note that the second derivative is the lowest order that is invariant upon reversal of (positive/negative) direction of the (space) axis, and therefore it does not discriminate between left and right. For this reason linear time processes involve first derivatives, linear space processes second derivatives. Concerning space metrics in discrete space models see the second chapter of this book (Commendatore et al. 2015).

Going from the line to true geographical two-space makes little difference. Continuing this heuristic kind of reasoning, the change of a function over the two-dimensional plane can be decomposed in a horizontal and a vertical component, and the two second derivatives can then just be added. The sum goes under the name Laplacian and measures how a variable changes over the plane when we leave a point—all possible directions of departure considered and combined to a net effect.

1.2.1 The Laplacian in Economics

The Laplacian appears in almost all classical systems of theoretical physics, such as heat diffusion, and mechanical waves in strings or membranes. (Of course physics also deals with processes in full 3-space, but in geography we never need to go beyond 2D.)

Only once was the Laplacian applied in economics, in Harold Hotelling's master's degree thesis Hotelling (1921) dating from 1923 which dealt with population growth and dispersal. Growth was modelled through a logistic Malthusian process, dispersal through linear diffusion away from densely to sparsely populated areas. The intuitive reason for such assumed dispersal offered was that with production under decreasing returns, living standard would be adversely related to population density, so people would move to more promising areas.

Hotelling's model was no success in economics—only few authors ever cited it, and it was not even included in the tiny volume of Hotelling's collected articles. Notably the same model was re-invented by Skellam thirty years later, applied to animal populations; as such it became the very basis of mathematical ecology, with thousands of followers.

Diffusion as used in Hotelling's model admittedly has a poor economics foundation, as the decreasing returns technology assumed and its relation to population (labour force) was never fully explained. The application of the Laplacian to population density in space was hence never firmly based on fundamental economics principles.

It can, however, with much better underpinning be used in models of spatial dispersion of economic growth and business cycles. Interregional trade models always used a linear import/export multiplier (see Metzler 1950), which can easily be extended to models involving continuous space, and then the linear Laplacian operator as a measure of spatial income differences is most appropriate.

2 Space Models of Economic Growth and Business Cycles

2.1 The Laplacian

Let us first introduce space in the lower order growth models. The Laplacian operator was verbally described above. We said there was a fundamental difference between time and space.

One is obvious: Geographical space is two-dimensional, time is one-dimensional. But there is a much more important and subtle difference: Time involves a forward/backward direction, which does not exist for left/right, even if space were one-dimensional.

This observation may seem trivial, but it has some unexpected consequences: A positive first time derivative $\frac{dz}{dt} > 0$ says that the variable z increases with time. In space, one wants to compare conditions in a point with conditions in *both* the surrounding directions (left *and* right). Moving from a point to the right, a *positive* derivative $\frac{dz}{dx}\Big|^{+} > 0$ says that z increases as one moves *away* from the point. Moving to the left a *negative* derivative $\frac{dz}{dx}\Big|^{-} < 0$ again says that z increases as one moves *away* from the point. To compare with both the surrounding "points", one needs both left and right derivatives, i.e., their difference (due to direction change) $\frac{dz}{dx}\Big|^{+} - \frac{dz}{dx}\Big|^{-}$. In the limit this difference becomes the second derivative $\frac{d^2z}{dx^2}$, if positive, z increases as one moves away from a point (is a concave function of x), if negative, z decreases (is a convex function of x).

Note that the second derivative is the lowest order invariant with respect to reversing directions, and that it enters as the lowest order space derivative in all models of linear processes in physics, such as diffusion, waves etc.

This also applies to higher dimension. In two dimensions the Laplacian operator $\nabla^2 z = \frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2}$ is the generalization. In one dimension an interval is just bounded by two points; in two an enclosed region is surrounded by a boundary curve of any irregularity. Gauss's Integral theorem states that the Laplacian is the proper measure of change in the variable z from the enclosure to its surroundings, no matter what shape the boundary curve has. For a simple explanation of this theorem, see Puu (1997/2003).

Formally

$$\iint_R \nabla^2 z\,(x_1, z_2)\; dx_1\, dx_2 = \oint_{\partial R} \nabla z\,(x_1, z_2) \cdot \mathbf{n}\, ds$$

Here R denotes a closed region of any shape, and ∂R its boundary curve. The shape is very general, the region need not even be simply connected, it can contain holes, provided they are accounted for in the boundary curve. In the left hand double integral we have the Laplacian of some function of the space coordinates which is integrated over the geographical region, on the right one integrates the gradient of this function projected on the boundary normal $\mathbf{n}$. Recall that the gradient is a directional derivative, whose projection on the outward normal tells us exactly the size of change (positive or negative) of the function $z\,(x_1, z_2)$ as we leave the region in any given boundary point. The line or curve integral on the right means that one integrates along the entire boundary curve, parameterized through $x_1\,(s)\,, x_2\,(s)$, always taking care to take a direction such that the interior of the region is to the left, and not forgetting possible holes.

The Gaussian theorem (sometimes also referred to as Green's theorem) is independent not only of the shape of the region, but also of its size. Hence, one can decrease its size around some chosen point to almost nothing, and then the theorem says that the Laplacian in the point can be interpreted as the net change of $z\,(x_1, z_2)$ as we leave that point, all directions of departure being taken in account.

2.2 Growth

Recall Harrod's model of balanced growth. It is based on two simple principles: (i) Consumption is a fixed proportion c of income, $C = cY$. Investment, due to the principle of acceleration is triggered by the rate of change of income, $I = a\frac{d}{dt}Y$. Investment is the change of capital stock, and, given a production technology of fixed proportions, capital stock in the proportion a to (real) income produced is needed. Using the accounting identity for a closed economy $Y = C + I$, one gets

$$a\frac{d}{dt}Y - (1 - c)\,Y = 0, \tag{1}$$

which has the obvious solution

$$Y = Y_0 e^{\frac{1-c}{a}t}.$$

Note that the model is formulated for a closed economy, which in a spatial setting for each point means no trade interaction whatever with its neighbourhood.

For an open economy the accounting identity becomes $Y = C + I + X - M$, where X and M denote exports and imports respectively. With Metzler (1950), assume a constant import propensity m. Then local income triggers imports, income in the surrounding neighbourhood triggers exports. Export surplus then equals this propensity m applied to the net spatial income change as one leaves the point, i.e., $X - M = m\nabla^2 Y$. Note that in some directions income may increase, in other decrease. Plugged into the accounting identity along with consumption and investment this gives

$$a\frac{\partial}{\partial t}Y - (1-c)\,Y + m\nabla^2 Y = 0. \tag{2}$$

2.2.1 Coordinate Separation

Considering a geographical region with some suitable boundary condition, such as prescribing zero growth on the boundary, it is natural to try a solution with separated coordinates, of the form

$$Y = T\,(t)\,S\,(x_1, x_2)\,.$$

Substituting this expression and its (time and space) derivatives in (2), and further dividing through by $T\,(t)\,S\,(x_1, x_2)$,

$$a\frac{1}{T\,(t)}\frac{d}{dt}T\,(t) - (1-c) + m\frac{1}{S\,(x_1, x_2)}\nabla^2 S\,(x_1, x_2) = 0. \tag{3}$$

Except for the constant, $(1-c)$, the equation contains one term, $a\frac{1}{T(t)}\frac{d}{dt}T\,(t)$, that only depends on time, another, $\frac{1}{S(x_1,x_2)}\nabla^2 S\,(x_1, x_2)$, that only depends on space coordinates. The first is hence independent of space, the second of time. In a solution where we want to cover time and space, both expressions must hence be regarded as constants.

Putting $\lambda = -\frac{1}{S(x_1,x_2)}\nabla^2 S\,(x_1, x_2)$, which can be rewritten

$$\nabla^2 S\,(x_1, x_2) + \lambda S\,(x_1, x_2) = 0. \tag{4}$$

we have a classical Eigenvalue/Eigenfunction problem for the *spatial* facts.

Then substituting $\lambda = -\frac{1}{S(x_1,x_2)}\nabla^2 S(x_1,x_2)$ in (3), we get $a\frac{1}{T(t)}\frac{d}{dt}T(t) - (1-c) - m\lambda = 0$, or multiplied through by $T(t)$,

$$a\frac{d}{dt}T(t) - (1 - c + m\lambda)\,T(t) = 0, \tag{5}$$

which is a pure *temporal* equation. It is linear and easily solved for each λ;

$$Ae^{\frac{1}{a}(1-c+m\lambda)t},$$

quite similar to the case of the original Harrod equation.

Now, there are infinite sequences of ascending Eigenvalues λ that solve the problem stated in (4), for each shape of the region and its boundary conditions, and each λ gives a different solution for (5)

If the region, for instance, is a unit square where all points of the boundary remain at rest at all times, then all Eigenvalues $\lambda = \sqrt{i^2 + j^2}$ with i, j any positive integers are solutions. The spatial patterns are rectangular subdivisions of the square at different scales, with regions of growth and recession alternating, though the combination of scaled rectangles can result in all but straight line boundaries.

The coefficients of the cosine and sine functions that solve (4) in any particular problem can be calculated through simple Fourier analysis from any initial spatial profile of the income distribution over the region. In addition to the boundary conditions we thus also need initial conditions for the solution.

Combining, the full solution reads

$$Y(t, x_1, x_2) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} A_{ij}\, e^{\frac{1}{v}(1-c+(i^2+j^2)m)t} \sin(\pi i x_1)\sin(\pi j x_2) \tag{6}$$

For calculating A_{ij} from the initial income profile take $t = 0$ in (6),

$$Y(0, x_1, x_2) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} A_{ij} \sin(\pi i x_1)\sin(\pi j x_2)$$

Integrating over the unit square one obtains

$$A_{ij} = \int_0^1\int_0^1 Y(0, x_1, x_2)\sin(\pi i x_1)\sin(\pi j x_2)\,dx_1 dx_2$$

See Puu (2000/2003), or any standard text on mathematical physics.

In the case of a circular region, we likewise deal with solutions in terms of Bessel Functions, and, with a heroic abstraction, considering the world economy on the entire globe, with Legendre Polynomials. See the present author's (Puu 2000/2003) for details.

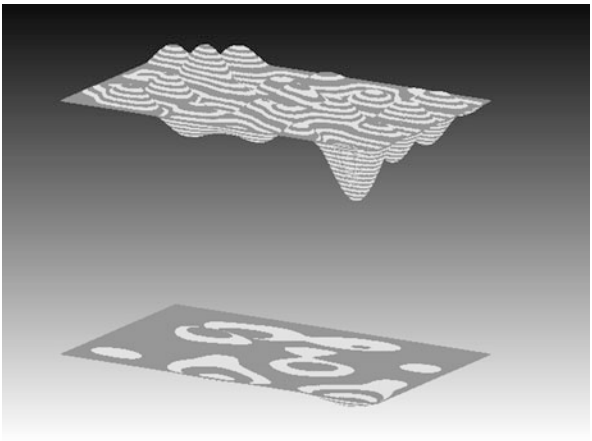


Fig. 1 A spatial pattern of change in the linear business cycle, and the projection of prosperity and depression areas on the horizontal plane at some given moment of time. The complex picture results from a mixture of different Eigenmodes which move at different speeds, thus resulting in different pictures at each frozen moment of time

Figure 1 illustrates an example of the distribution of growth and decline on a square region. The same figure can, by the way, also illustrate a linear business cycle on a square region in progress at a frozen time moment.

The temporal change profiles that solve (5) are simple exponentials as in the original Harrod model; as a rule growth is faster the smaller the spatial subdivision. Like all linear models also the present only has meaning for short time periods, as the exponentials either extinguish motion, or else make the solution explode. This also holds for the case of linear business cycles dealt with next. For sustained and bounded motion one needs nonlinear models, but these, as a rule, are almost impossible to handle without the use of numerics. Efficient numerical methods are also needed for handling more complicated regions than those mentioned. As a rule these request triangulation of the plane region. One may think that in analogy to discretization of the line in intervals, division of the plane in square cells might work, but this is not true.

Further, the idea of division in square cells, and dynamics through contamination through contiguous cells may lead one to think of the popular "game of life", but one should be clear about the fact that such analogies are totally meaningless for any problems of regional economics as they assume a torus shape of the region. Whatever the shape we want to deal with, we can be sure that there never existed any geographical region shaped like a torus.

2.3 Business Cycles

2.3.1 The Linear Phillips Model

Through increasing the order of the differential equation in Harrod's growth model, oscillating systems to produce business cycles could be obtained. This was achieved by Phillips (1954) and Allen (1956). Harrod was fully aware of the fact that his

balanced growth path was an unstable solution, and that any initial deviation from it would lead to paths that diverged from it, but he failed to model what would happen at such deviations. This was left to Phillips who used adaptive delays.

Suppose income increases in proportion to the excess of planned expenditures (consumption plus investment) over actual income,

$$\frac{d}{dt}Y(t) = C(t) + I(t) - Y(t), \tag{7}$$

which now replaces the accounting identity. Through a change of scale for time or income proportionality in the right hand side can be replaced by equality, thus saving us from dragging along an unnecessary adaptation coefficient. As for consumption, let still

$$C(t) = cY(t), \tag{8}$$

but assume that investment, due to the delay in constructing capital equipment, is just adjusted in the proportion to which the accelerator triggered optimal investment $a\frac{d}{dt}Y$ exceeds or falls short of actual investment, i.e.,

$$\frac{d}{dt}I(t) = a\frac{d}{dt}Y(t) - I(t). \tag{9}$$

Again a change of scale can rid us of another adaptation coefficient.

Differentiating (7) once more

$$\frac{d^2}{dt^2}Y(t) = \frac{d}{dt}C(t) + \frac{d}{dt}I(t) - \frac{d}{dt}Y(t).$$

Next, differentiation of (8) yields $\frac{d}{dt}C = c\frac{d}{dt}Y$, which substituted along with (9) in the previous equation yields

$$\frac{d^2}{dt^2}Y(t) = (a + c - 1)\frac{d}{dt}Y(t) - I(t).$$

Finally, from (7) and (8) $I = \frac{d}{dt}Y + Y - C = \frac{d}{dt}Y + (1 - c)Y$, so,

$$\frac{d^2}{dt^2}Y(t) = (a + c - 2)\frac{d}{dt}Y(t) - (1 - c)Y(t). \tag{10}$$

The differential equation is now second order and can thus also produce oscillations, which is Phillips's version of Samuelson's multiplier-accelerator model. Note that in continuous time adaptation works quite as distributed time lags do in discrete time. Also note that Phillips's model is constructed along economics principles and that the relation between the discrete and continuous models does not involve playing any mathematical tricks such as using the first recurrence map

on a Poincaré section, or any standard recipe for discretizing a continuous system from the numerical analysis repertory.

2.3.2 Trade and the Open Economy

Making the economy open, and introducing exports and imports, we can use the same adaptation scheme for export surplus as for investment. Then (10) changes to

$$\frac{\partial^2}{\partial t^2} Y(t, x_1, x_2) - (a + c - 2) \frac{\partial}{\partial t} Y(t, x_1, x_2) + (1 - c) Y(t, x_1, x_2) - m \nabla^2 Y(t, x_1, x_2) = 0 \tag{11}$$

Again it is possible to separate temporal and spatial processes, now to obtain spatial waves. Putting $Y(t, x_1, x_2) = T(t) S(x_1, x_2)$ one again gets the Eigenvalue/Eigenfunction problem

$$\nabla^2 S(x_1, x_2) + \lambda S(x_1, x_2) = 0$$

(4) back, whereas the temporal equation (5) changes to

$$\frac{d^2}{dt^2} T(t) - (a + c - 2) \frac{d}{dt} T(t) + (1 - c + m\lambda) T(t) = 0. \tag{12}$$

The spatial patterns are the same as in the case of growth; they depend on the kind of region and boundary conditions we choose. Of course, regular shapes such as the square or the disk, provide nice closed form solutions in terms on trigonometric or Bessel functions. The same holds for the sphere where Legendre polynomial can be used. For mathematical detail, see any of the present author's books, Puu (1997/2003) or Puu (2000/2003). Irregular shapes are, naturally, more difficult to deal with.

Each solution to (4) provides an eigenvalue λ, which substituted in (12) results in a different second order ordinary differential equation with an oscillatory solution where areas of prosperity are separated by node lines from areas of depression, but interchange their phase in the cycle. As the system is linear, all the solutions for different λ can be combined, smaller subdivisions oscillating faster. The coefficients can again be calculated using Fourier analysis, though, due to the higher order we now need two initial conditions in terms of the initial income distribution and its velocity profile.

2.4 Nonlinear Business Cycle Models

The main drawback of linear models is that the motion either becomes extinct or explodes. In 1989 the present author suggested a non-linear accelerator to the Phillips model, including a negative third order term to the purpose of accounting for the Hicksian floor and ceiling constraints. See Puu (1989). Without including spatial

diffusion, the model resulted in a Rayleigh/van der Pol type of nonlinear oscillator capable of producing bounded but persistent motion. See Stoker (1950), van der Pol (1926), Hayashi (1964). The resulting model was studied by the Poincaré–Lindstedt perturbation method.

Through coupling, again using a linear trade multiplier, though applied to two point economies, or just forcing, to mimic one large and one small economy, indications of possible chaotic motion were obtained. Considering continuous space with infinitely many contiguous locations, things are bound to become much more complicated.

Let us present the spatial model proposed in Puu (1989), and indicate just one possible type of solution for it

$$\begin{aligned} \tfrac{\partial^2}{\partial t^2} Y\left(t, x_1, x_2\right) + (1-c)\, Y\left(t, x_1, x_2\right) - m \nabla^2 Y\left(t, x_1, x_2\right) = \\ (a+c-2)\, \tfrac{\partial}{\partial t} Y\left(t, x_1, x_2\right) - a\left(\tfrac{\partial}{\partial t} Y\left(t, x_1, x_2\right)\right)^3 \end{aligned} \tag{13}$$

Notably it is no longer possible to separate spatial and temporal parts of the solution, neither can solutions be combined.

As we said, we can just illustrate the very rich solution family by one case. Assume the wave surfaces are flat; then the Laplacian, the sum of the second derivatives $\frac{\partial^2 Y}{\partial x_1^2} + \frac{\partial^2 Y}{\partial x_2^2}$, is zero, i.e. $\nabla^2 Y\left(t, x_1, x_2\right) = 0$. By the way, flat surfaces are not the only case with a zero Laplacian—a saddle shaped surface has the same property, as has any solution to the Laplacian differential equation which reads just $\nabla^2 Y\left(t, x_1, x_2\right) = 0$. Whether we can use saddles or flat surfaces all depends on the boundary and boundary conditions. This illustrates that boundary conditions in partial differential equations are indeed an integral part of the problem, and that the form of the equation itself, such as (13) on its own is incomplete and hence cannot be solved.

Further, assume the growth rate of income takes any of the three constant values, $\frac{\partial}{\partial t} Y = \pm\sqrt{\frac{a+c-2}{a}}$, or 0. Accordingly, the right hand side of (13) as well is zero. Further, as $\frac{\partial}{\partial t} Y$ is constant in all three cases, the second time derivative too becomes zero, i.e., $\frac{\partial^2}{\partial t^2} Y = 0$.

The result is a wave front $\alpha x_1 + \beta x_2$ moving with the constant speed $\pm\sqrt{\frac{a+c-2}{a}}$ or 0, i.e.,

$$Y\left(t, x_1, x_2\right) = \alpha x_1 + \beta x_2 \pm \begin{cases} \sqrt{\frac{a+c-2}{a}} \\ 0 \\ -\sqrt{\frac{a+c-2}{a}} \end{cases} t, \tag{14}$$

Introducing (14) in (13), one gets

$$(1-c)\, Y\left(t, x_1, x_2\right) = 0,$$

which implies $Y\left(t, x_1, x_2\right) = 0$.

This combination; flat surfaces of points moving up or down at constant speeds or standing at rest obviously fulfill Eq. (13). It is only needed to piece everything together to a consistent global picture. If we have a square region, say $|x_1| \leq 1 \wedge |x_2| \leq 1$, and boundary conditions that prescribe constant rest at the edges, then the solution wave takes the form of a pyramid which is shaved flat when time passes, and, after becoming erased flat to the ground, is then inverted into negative reflections in the horizontal plane of the pyramids; and so it goes for ever. Formally, at each moment of time only the square

$$x_1, x_2 = \pm\sqrt{\frac{a+c-2}{a}}t,$$

and its interior is moving at the speed $\sqrt{\frac{a+c-2}{a}}$ or $-\sqrt{\frac{a+c-2}{a}}$, whereas all the remaining points, on the sides of the pyramid stay at rest. This can go on until the top of the pyramid right above the origin, or its bottom below is reached; when, obviously, the speed is reversed in sign (Fig. 2).

No doubt, this is a solution. The question is if it is attracting. Given the square shape and the boundary conditions, both Poincaré–Lindstedt perturbation analysis, and numerical study indicated that the system is attracted to a sustained bounded cycle where the shapes take the form of pyramids. Again see Puu (2000/2003)

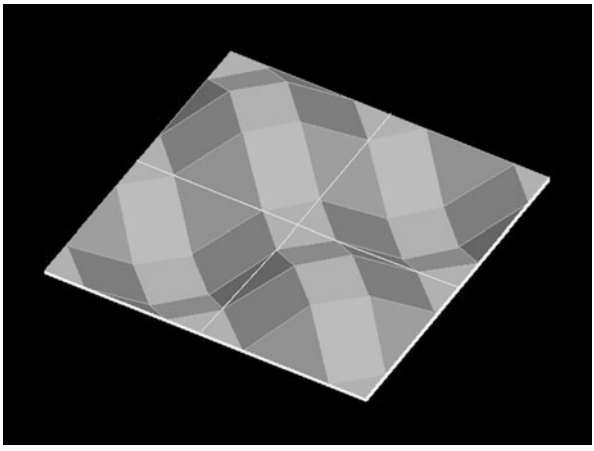


Fig. 2 A pattern of four shaved pyramids that solve the nonlinear business cycle model. Note that only the flat squares on the *tops* and *bottoms* of the truncated pyramids move up or down, whereas the walls of the pyramids and the horizontal surfaces stay still. In the movement these horizontal surfaces thus decrease or increase in size. In the former case the movement goes on until the surface shrinks to a point so that the entire pyramids and their excavated mirror images are complete, in the latter case the pyramids are eventually completely levelled with the ground. In these extreme situations, movement speed must obviously shift to another of the three possible speeds. Note that this is but one possible solution scenario, due to the special shape of the region and the boundary conditions imposed

for detail. Different initial wave conditions, depending on their mesh size produce different solutions of the same pyramid shape but with internal node lines. However, in a nonlinear system these cannot be combined as they can in the solution to the wave equation.

2.5 Boundaries and Boundary Conditions

2.5.1 Region Shapes

In all cases of modelling with partial differential equations, the definition of boundaries and boundary conditions is as important as formulating the equation itself. In the examples above we chose simple boundaries, such as squares, and constancy (equilibrium) as condition along the entire boundary. This sets the stage for standing waves.

In space without a boundary, such as the infinite plane or the surface of a sphere (topologically equivalent through the stereographic projection if we remove the north pole from the sphere) the stage is instead set for travelling waves. However, the sphere is almost the only case of such a region which can make sense in a geographical context (as a picture of the global world economy).

2.5.2 Boundary Conditions

Interesting regions have boundaries, though of any shape. Whenever we study an aggregate of adjacent regions, it is interesting to find out how impulses are propagated through boundaries, with or without obstacles put on the passage of cyclic waves. Therefore it is interesting to study the sensitivity of, say, a region within which cycles are damped, so that left to itself it would go to rest, when excited from the exterior through periodic boundary conditions. What importance do size and shape have on resilience to such exterior disturbances, and what resonance relation is there between the exciting disturbance frequency and the Eigenvalues of the excited region? Do complex dynamic phenomena arise if a non-damped regional economy in stead is excited in this way?

2.5.3 Software

It must be realized that most region shapes, with the rare exception of those developed with classical mechanics, must be dealt with numerically. Unlike the case of most physical phenomena which are three-dimensional, our topic is confined to two dimensions. This gives unique opportunities for visualization; as for instance income fluctuations over systems of geographical regions become fluctuating surfaces that can always be displayed in projection. It therefore is of primary importance to check what existing software can offer and to complement it with new software if necessary.

References

Allen, R. G. D. (1956). *Mathematical economics*. London: MacMillan.

Beckmann, M. J. (1952). A continuous model of transportation. *Econometrica, 20*, 642–660.

Commendatore, P., Filoso, V., Grafeneder-Weissteiner, & T., Kubin, I. (2015). Towards a multiregional NEG framework: Comparing alternative modelling strategies. In P. Commendato, S. Kayam, & I. Kubin (Eds.), *Complexity and geographical economics: Topics and tools*. Heidelberg: Springer (this volume). doi:10.1007/978-3-319-12805-4_2.

Duesenberry, J. (1950). Hicks on the trade cycle. *The Quarterly Journal of Economics, 64*, 464–476.

Frisch, R. (1933). *"Propagation problems and impulse problems in dynamic economics". Economic essays in honour of Gustav Cassel*. London: Allen & Unwin.

Hansen, A. H. (1951). *Business cycles and national income*. New York: Norton.

Harrod, R. F. (1948). *Towards a dynamic economics*. London: MacMillan.

Hayashi, C. (1964). *Nonlinear oscillations in physical systems*. Princeton, NJ: Princeton University Press.

Hicks J. R. (1950). *A contribution to the theory of the trade cycle*. Oxford: Oxford University Press.

Hotelling, H. (1921/1978). *A mathematical theory of migration*. MA Thesis, University of Washington. Reprinted in *Environment and Planning A, 10*, 1223–1239.

Keynes, J. M. (1936). *The general theory of employment, interest, and money*. London: MacMillan.

Lindahl, E. (1939). *Studies in the theory of money and capital*. New York: Farrar & Rinehart.

Metzler, L. A. (1950). A multiple-region theory of income and trade. *Econometrica, 18*, 329–354.

Phillips, A. W. (1954). Stabilisation policy in a closed economy. *The Economic Journal, 64*, 290–323.

Puu, T. (1989). *Nonlinear economic fynamics, lecture notes in economics and mathematical systems* (Vol. 336). Berlin: Springer.

Puu, T. (1997/2003). *Mathematical location and land use theory*. Berlin: Springer.

Puu, T. (2000/2003). *Attractors, bifurcations, & chaos – nonlinear phenomena in economics*. Berlin: Springer.

Puu, T., & Sushko, I. (Eds.) (2006). *Business cycle dynamics*. Berlin: Springer.

Samuelson, P. A. (1939) Interactions between the multiplier analysis and the principle of acceleration. *Review of Economics and Statistics, 21*, 75–78.

Stoker, J. J. (1950). *Nonlinear vibrations in mechanical and electrical systems*. New York: Wiley.

van der Pol, B. (1926). On relaxation oscillations. *Philosophical Magazine, 2*, 978–992.

von Haberler, G. (1937). *Prosperity and depression*. Cambridge, MA: Harvard University Press.

Part II
Institutions and Markets

Complex Financial Networks and Systemic Risk: A Review

Spiros Bougheas and Alan Kirman

Abstract In this paper we review recent advances in financial economics in relation to the measurement of systemic risk. We start by reviewing studies that apply traditional measures of risk to financial institutions. However, the main focus of the review is on studies that use network analysis paying special attention to those that apply complex analysis techniques. Applications of these techniques for the analysis and pricing of systemic risk has already provided significant benefits at least at the conceptual level but it also looks very promising from a practical point of view.

1 Introduction

The 2007–2009 global financial crisis has painfully demonstrated how costly a systemic failure can be. Systemic events impose high costs on taxpayers as governments usually intervene by bailing out important institutions in an effort to ensure the survival of the financial system. Unfortunately, the implementation of such policies are also part of the causes of the next crisis. This is because they encourage opportunistic behavior by these same institutions in anticipation of the government policies. This "moral hazard" problem is central to understanding the dilemma with which policy makers are faced.

The study of financial crises has been on the research agenda of financial economists for a long time now.[1] One particular aspect of systemic crises that has

[1] See Brunnermeier and Oehmke (2012) for a review of this literature.

S. Bougheas (✉)
School of Economics, University of Nottingham, Nottingham NG7 2RD, UK
e-mail: spiros.bougheas@nottingham.ac.uk

A. Kirman
Faculte de Droit et de Sciences Politiques, Aix-Marseille Université, Directeur d'études à l'EHESS, Membre de l'IUF, GREQAM, Bureau 109, 3 avenue Robert-Schuman, 13628 Aix-en-Provence, France
e-mail: alan.KIRMAN@univmed.fr

P. Commendatore et al. (eds.), *Complexity and Geographical Economics*,
Dynamic Modeling and Econometrics in Economics and Finance 19,
DOI 10.1007/978-3-319-12805-4_6

recently attracted a lot of attention is that of contagion. Even if an initial shock affects only a small number of institutions, the connectivity of the financial system (for example, because of the interbank market, the payments system, portfolio correlations. etc.) implies that the shock will be transmitted widely, and will increasingly often cross international boundaries.[2] Some researchers have applied the tools of network theory, that is ideally suitable for the analysis of interconnected systems, to the study of systemic events.[3] What becomes clear from this work is that both the number of affected institutions and the financial system's volume of losses depend not only on the aggregate volume of risk exposures but also on their distribution within the system and the structure of that system.[4]

Simple models are very useful for providing a conceptual framework so that we can gain an intuitive understanding of the relationship between network structure and systemic risk. However, they fall short of providing practical guidance to regulators and policymakers in relation to actual financial networks that consist of a large number on institutions with a huge number of links between them. What we need are tools for the study of such systems so that we can get better estimates of both the likelihood of contagion and systemic risk. The ecologist Robert May has been a strong advocate of the application of complex systems analysis to the study of financial networks.[5] Over the last 10 years, many researchers have followed this promising line of research by using these analytical tools to improve our understanding of the financial sector in general and banking systems in particular.[6] In this paper, we review this new and fast growing literature paying special attention to issues related to the measurement of systemic risk. In the following section we review various proposals for measuring systemic risk that use traditional finance tools. In the subsequent two sections we review theoretical and empirical contributions that employ complex networks techniques to the analysis of systemic risk. In the final section we briefly evaluate the progress that has been made so far and identify areas for further research.

[2]For an interesting exposition of the geographic distribution of bank failures during the recent global financial crisis see Aubuchon and Wheelock (2010).

[3]For expositions of network theory and its applications to economics see Goyal (2009), Jackson (2008) and Vega-Redondo (2007).

[4]See Allen and Babus (2009) for a review of various applications of network theory to the study of financial issues.

[5]See, for example, Haldane and May (2011), May et al. (2008). For a similar perspective on the benefits of network analysis, see also Schweitzer et al. (2009).

[6]There is an established literature, known as econophysics, that has used complex systems to analyze various economic systems including the behavior of asset prices (for a review see Varela et al. 2015, this volume).

2 Non-network Approaches to Measuring Systemic Risk

Since the 2007–2009 global financial great emphasis has been placed on developing both theoretical and practical measures of systemic risk.[7] Some researchers have opted for a structural approach while others have worked with reduced-form approaches that focus on the behavior of asset returns in the tail of the distribution. The structural approach requires a general equilibrium model and explicit restrictions on structural errors so that the parameters of behavioral functions can be estimated. In contrast, the goal of reduced-form studies is to provide estimates for these parameters using exogenous within-sample variation minimizing the reliance on structural assumptions.

The structural approach is followed by Gray et al. (2008) who propose the use of Contingent Claims Analysis (CCA) to evaluate the sensitivity of an economic system's balance sheets to external shocks. CCA treats risky debt, which raises the possibility of default and thus systemic risk, as a put option.[8] If one type of risky debt (loans from banks to firms) is linked to another type of risky debt (loans from banks to banks), the second type of risky debt can be expressed as a function of the implicit put option of the first type of risky debt. Strong non-linearities in risk transmission can potentially arise from this compound nature of interlinked risky debt.[9]

Segoviano and Goodhart (2009) view the banking system as a portfolio of banks from which they infer the banking system's multivariate density. Their measures reflect not only linear dependencies within the system (correlations) but also non-linear distress dependencies that alter through time. Employing a non-parametric copula approach they are able to distinguish between (a) common distress in the system, (b) distress between specific banks, and (c) cascade effects. They use their approach to derive a variety of bank stability measures and then use them to estimate changes in the stability of the US banking system over time, to estimate cross-regional effects between American and European banking groups and to assess the impact of international banks on sovereigns.

One advantage of the structural approach is that it allows one to derive consistent measures for the marginal risk contribution of each institution in the system. These measures are consistent in the sense that they add up to the aggregate systemic risk. This has been the motivation behind the work by Tarashev et al. (2010) who use a game-theoretic methodology to apportion system-wide risk to individual institutions. In particular, they use the Shapley Value (a solution for

[7] For a review of systemic risk measures applied to general financial systems see Markellof et al. (2012). See also European Central Bank (2007) for some theoretical and empirical contributions related to the measurement of systemic risk in banking and non-banking financial systems.

[8] A put option is a derivative that offers the right, but not the obligation, to the holder to sell the underlying asset at a pre-specified price within a given period.

[9] For an example of CCA see Lehar (2005) who applies the methodology to a sample of international banks from 1988 until 2002.

cooperative games) that allocates payoffs to players that are equal to their marginal contributions. The idea is simple but appealing. The marginal contribution to risk of an institution to a group of banks is the amount by which the aggregate risk of the group increases when that bank is added. Consider all the possible groups of banks and evaluate the marginal contribution of the particular bank to each of them. The overall contribution to risk of the bank in question is then a weighed sum of all these marginal contributions. Huang et al. (2012) also construct systemic risk measures satisfying consistency, but use a different approach. They use equity-price comovements and Credit Default Swap (CDS)[10] spreads to construct hypothetical insurance premiums against catastrophic losses in the banking system. The first factor provides a measure of asset return correlations while the second measures the probability of default. Applying their methodology to the US banking system the authors find that a bank's systemic risk contribution is primarily determined by its size and to a lesser extent by the correlation of asset returns and default probability.

CDS prices are also used, together with bond prices, by Giglio (2011) who develops an alternative method for measuring joint default risk of large financial institutions. Bond prices offer information about the default probability of the issuer while CDS prices, that pay only if the seller of the insurance contract is solvent, contain information about the probability of joint default. This allows one to derive pairwise default probabilities within the financial system. However, it is not sufficient to characterize the systemic risk (joint default risk) of multiple institutions. Then, the author develops a general theory of probability bounds that permits the construction of tightest bounds for probabilities of high-order events given the availability of a low-order information set. The author applies the methodology to follow the evolution of default risk of large banks during the 2007–2009 financial crisis and finds that the majority of spikes in bond and CDS prices corresponded to changes in idiosyncratic rather than systemic risk.

Turning now to the studies that focus on the tail of the return distribution, Acharya et al. (2010) derive a theoretical measure of systemic risk, namely the Systemic Expected Shortfall (SES) which is equal to the expected amount that a bank is undercapitalized in a future systemic event in which the overall financial system is undercapitalized. Their theory suggests that the regulation of systemic risk should depend on each institution's SES and the overall probability of a systemic event. For practical purposes regulators need to estimate the conditional expected losses before a crisis occurs and according to the theory they should use any variable that is a good predictor of capital shortfall during a systemic event. They propose that the regulators estimate the Marginal Expected Shortfall (MES) that corresponds to the amount of expected losses during moderately bad days together with the level of leverage as predictors for SES. The authors use data from 2007–2009 during the financial crisis to demonstrate that their measure of systemic risk would have been

[10]A CDS is a financial swap agreement whereby the seller of the contract will compensate the buyer in the event of a loan default. The buyer of the CDS makes a series of payments to the seller but only receives a payoff if the loan defaults.

a good predictor of subsequent measures of risks, such as outcomes of stress tests, the decline in equity evaluations, and the widening of credit default swap spreads.

Hartmann et al. (2005) apply Extreme Value Theory (EVT) on bank equity prices which allows them to estimate the probability of contagion between banks, the sensitivity of banks asset returns to aggregate shocks and also the fluctuations of those risks over time. In particular, they construct two indicators of systemic risk. The first indicator captures bank contagion risk by measuring extreme comovements of bank asset returns. The second indicator captures the impact of aggregate portfolio shocks on individual banks' stock returns. The authors apply their methodology to US and European banking systems finding that while systemic risk has increased in both regions during the 1990s, systemic risk is higher in the US than in the Euro area, probably due to the weaker cross-border European banking linkages relative to U.S. ones.[11]

Value at risk (VaR) is a common measure of risk used by financial institutions. For example the 5 %-VaR is the maximum loss within the 5 % confidence interval. Clearly, a single institution's measure of risk cannot reflect systemic aspects of risk. Adrian and Brunnermeier (2011) propose an alternative measure that they refer to as CoVaR. An institution's CoVaR is equal to the VaR of the whole system conditional on that institution being in a particular state (distress or no-distress). By taking the difference between the CoVaR conditional on the institution being in distress minus the CoVaR conditional on the institution not being in distress the authors derive ΔCoVaR that captures the marginal contribution of that particular institution to systemic risk. However, from a regulatory point of view what is needed is a forward looking measure that would be a good predictor of future systemic events. To this end, the authors estimate forward looking time-varying measures of ΔCoVaR using weekly data from 1986Q1 to 2010Q4 and show that the 2006Q4 value would have predicted over half of realized covariances during the global financial crisis.[12]

More closely related to the network literature, Billio et al. (2012) derive econometric measures of the degree of connectedness, and hence systemic risk, for four financial sectors, namely, hedge funds, banks, brokers, and insurance companies. They use principal-components analysis to uncover the degree of connectedness among the institutions in the four sectors and then they apply Granger causality tests in order to identify the direction of links. Their results suggest that among the four sectors the most important in transmitting shocks is the banking sector.

[11] The second indicator, known as Tail-β, has been also applied to the European banking system by De Jonghe (2010).

[12] Applications of this methodology include Wong and Fong (2010) who provide CoVaR estimates for the CDS of Asia-Pacific Banks and Gauthier et al. (2009) who give systemic risk estimates for the Canadian banking system.

3 Systemic Risk in Complex Financial Networks: Theoretical Developments

There is a steadily increasing literature suggesting that our understanding of systemic risk, and hence our ability to measure it, will be greatly enhanced by taking a close look at the topological properties of the network of transactions that link financial institutions together.[13] A typical financial network can be graphically represented by a set of nodes, one for each institution, and a set of weighted and directed links representing the bilateral financial relationships between those institutions. Depending on the application these values can measure either financial exposures (e.g. interbank deposits) or financial transactions (e.g. interbank payments market) over a period of time. Figure 1 shows the graph of a four-bank network where a link connecting bank i with bank j and with the arrow pointing to bank j represents the deposits of bank i at bank j. The thicker the arrow, the higher the weight of the link, that is the level of deposits. Thus, the graph captures the level of bilateral exposures.

The graph of a network of n banks can also be presented by a $n \times n$ real-valued matrix A. An entry α_{ij} represents the level of deposits of bank i at bank j. For example, a matrix corresponding to the network shown in Fig. 1 is given by

$$A = \begin{pmatrix} 0 & 100 & 50 & 0 \\ 50 & 0 & 0 & 100 \\ 0 & 0 & 0 & 50 \\ 100 & 0 & 0 & 0 \end{pmatrix}$$

Notice that along the diagonal all entries are equal to 0 capturing the fact that a bank does not hold deposits with itself. To keep things simple we have considered only three possible levels of exposure. In the above matrix entry $\alpha_{12} = 100$ which corresponds, in Fig. 1, to the high weighted arrow connecting bank 1 with bank 2

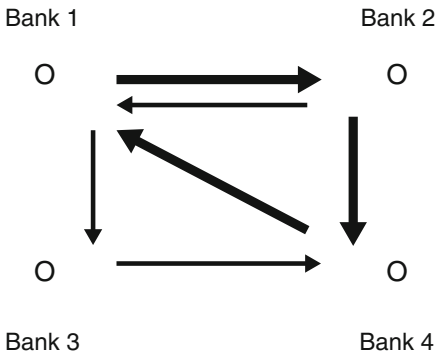


Fig. 1 Four-bank network

[13] For a more thorough exposition of network theory see Varela et al. (2015) in this volume.

and pointing at bank 2 and indicating that bank 1's level of deposits at bank 2 is equal to 100.

As we explain in more detail below, introducing a shock into the system is equivalent to disturbing some of these links. Analyzing the systemic consequences of that shock involves not only tracing the path of links that connects the institution where the shock was realized with the rest of the system but also taking into consideration their weights. In what follows we focus on research related to the analysis of banking systems, starting with simple networks and then moving to complex ones.[14]

3.1 Simple Banking Networks

This literature builds on some earlier work on interbank relationships that explores the negative externality that a failure of one bank imposes on the rest of the system; see, in particular, Bitzer and DeMarzo (1992) and Rochet and Tirole (1996). The main contribution of the following papers is the recognition that in a multi-banking system the specific link pattern of the network matters.

For example, Allen and Gale (2000) analyze the transmission of liquidity shocks within a four-bank network where the role of each bank is to provide liquidity insurance as in the Diamond and Dybvig (1983) and Allen and Gale (1998).[15] The authors find that the systemic impact of an initial shock crucially depends on the structure of the network. When all banks are interconnected to each other, that is the structure of the network is complete, then cross-holding of liquidity can be sufficient to absorb the shock without banks having to liquidate long-term assets. In contrast, when the structure is incomplete, the impact of the shock might be felt very strongly by banks immediately linked to the one that suffered the initial shock resulting in the liquidation of their long-term assets. As each subsequent link is affected the crisis spreads throughout the system.

Allen and Babus (2010) compare the systemic risk consequences of two network structures each consisting of six banks. In both structures each bank forms links with two other banks. In one structure, which the authors call clustered, there are two networks each consisting of three banks. In the unclustered version there is a single network with a cyclical structure. The authors concentrate their analysis on the case of short-term debt that requires rollover between periods and hence raises the possibility of systemic risk. For risk diversification purposes banks exchange

[14]Other researchers use complex analysis for understanding interfirm financial (trade-credit chains) networks; see Battiston et al. (2007), Cossin and Schellhorn (2007), Kiyotaki and Moore (2004) and Lagunoff and Schreft (2001).

[15]The two-bank version of the Diamond and Dybvig (1983) is analyzed by Bhattacharya and Gale (1987), however, a minimum of three banks are needed for comparisons of network structures to become meaningful.

assets within their network. The exchange of assets implies that under the clustered structure banks hold identical portfolios while this is not the case in the unclustered version. The authors find that the likelihood of early liquidation and thus systemic risk is higher in the clustered version but, if the proceeds from early liquidation are sufficiently high, the welfare of depositors can still be greater under the clustered structure.

Leitner (2005) takes one step back by asking if having a network offers any advantages. The author compares a complete network with one where every node is isolated and finds a trade-off. On one hand, there are insurance benefits due to the cross-links between banks but, on the other hand, this also raises the possibility of the whole network collapsing. Furthermore, when liquidity is concentrated in a small number of institutions the latter might opt not to participate. Similarly, Castiliognesi and Navarro (2007) find that the network structure depends on the level of counterparty risk. The network is generally characterized by a core-periphery structure where the core consists of safe banks that are fully connected but the degree of connectivity between the core and the periphery (risky banks) depends on the level of counterparty risk. Issues related to network formation are also addressed. In this model network formation is endogenous and the author shows that the optimal network structure is stable and reduces the risk of contagion.

3.2 Complex Banking Networks

A bank or a group of banks, can be potentially affected by two types of shocks. A sudden strong demand for liquidity can force the banks, after they exhaust their reserves, to use their deposits at other banks. This can initiate a domino effect that, depending on the network structure, can spread throughout the system. As long as there are enough liquid assets in the system, interbank deposits might be sufficient to absorb the demand for liquidity. However, if for some banks their deposits in the rest of the system are not sufficient to meet the demand for liquidity they might have to liquidate some of their long-term assets. If selling those assets depresses their prices this can lead to insolvency which is related to the second type of shock that is a direct hit on the asset side of the balance-sheet. In this case a bank has to write-off a significant fraction of its long-term assets mainly caused by the inability of its borrowers to pay back their debts. The theoretical literature that uses complex network techniques to address issues of contagion focuses on the systemic risk consequences of the second type of shock.

Caballero and Simsek (2009) analyse information transmission in a cyclical network structure where each bank is directly informed about its risk exposure to its neighbors. While most of the time this structure is sufficiently stable, in periods of high financial distress, when each bank becomes interested in its exposure to other banks through indirect links, the cost of information gathering becomes unmanageable and an financial crisis ensues. In a related study Anand et al.

(2011) use a mean-field approximation to analyse how banks can coordinate on a strategy which can lead to inefficient information aggregation and can then switch to coordinate on another strategy in which all information is revealed and the market crashes.

Many authors examine how network connectivity is related to systemic risk. Iori et al. (2006) find that when the network of banks is heterogenous, an interbank market can have ambiguous consequences for systemic risk. The benefits of mutual insurance need to be contrasted to the possibility of contagion. The authors also report that as the connectivity increases the system becomes more stable, echoing the finding of Allen and Gale (1998). Similar conclusions in relation to the degree of network connectivity are reached by Gai and Kapadia (2010), Montagna and Lux (2013) and Nier et al. (2007). The former paper also compares the impacts of indiosyncratic and aggregate shocks. With relatively high capital to total assets ratios the system can survive the former type of shock but with aggregate shocks contagion risk significantly increases. Montagna and Lux (2013) analyse scale-free networks while the other papers focus on random networks.[16] A scale-free topology arises endogenously in Da Cruz and Lind (2012). The authors show that the distribution of 'avalanches' (systemic events) follows a power-law. In their model higher capital requirements by offering incentives to banks to increase the market concentration of the banking system can enhance the likelihood of systemic events.

Two related topics that have attracted a lot of attention in the finance literature as a result of the global financial crisis are 'fire sales' and 'market freezes'.[17] Both of these issues have also been addressed using the network approach. For example, Cifuentes et al. (2005) analyse the transmission of shocks in systems that use marking to market practices. In such systems, sales of assets by depressing asset prices can induce further round of sales. The authors show that small initial shocks can generate huge systemic effects and that the exact aggregate exposure depends on the particular structure of the network. They also suggest that imposing liquidity requirements can be more effective than capital regulations in averting a crisis. May and Arinaminpathy (2010) use a mean-field approach to derive analytical results in relation to fire sales. They demonstrate how the impact of contagion depends not only on the size of the shock and the level of asset devaluation but also on the asset classes affected by the shock. Nier et al. (2007) observe that the effects of fire sales are more pronounced in higher concentrated systems. An example of a market freeze where banks decline to rollover debt is offered in Anand et al. (2012). The authors find that the spread of contagion depends on the maturity structure of loans and the rate at which adverse news spread in the system. When the endogenous probability

[16]Random networks correspond to the classical random graphs structures introduced by Erdös and Rényi (1959).

[17]For traditional finance models addressing these issues, see Acharya et al. (2011) and Diamond and Rajan (2011).

of bank failure is sufficiently high their model also generates market freezes where banks decline to lend to each other.

Some researchers compare the impact on systemic risk of various network structures. Acemoglu et al. (2013) compare complete networks with networks that have a ring structure. They find that although complete networks are generally more stable, under extreme conditions,[18] the high number of interconnections can also be responsible for higher fragility. The authors also assess the welfare properties of network systems formed in a decentralized way and they find that, as banks do not internalize the negative externality that their liquidity management decisions impose on the rest of the system, overall aggregate liquidity is suboptimal.[19] Contagion and cascades are analysed using variety of network structures by Elliott et al. (2013). In particular, the authors explore how changes in the degrees of integration (how far an institution depends on its counterparties) and diversification (number of counterparties) affect the spread of defaults through the system. Lenzu and Tedeschi (2012) compare systemic risk in random formed and scale-free networks. The authors find that scale-free networks are more vulnerable because of (a) sub-optimal liquidity allocation, and (b) higher heterogeneity among participants which increases the exposure of the banking system to contagion. Teteryatnikova (2012) assesses the advantages of multi-tier systems finding that the resilience of the banking network to systemic shocks increases with the level of tiering.

In an alternative application of network analysis Eisenberg and Noe (2001) develop an algorithm that not only provides an efficient way for clearing the interbank network following a shock but also provides measures of the systemic risk by each bank in the system. Their results suggest that an increase in the system's cash flow volatility by reducing the payments across institutions lowers the asset value of the banking system. Lastly, more recently, Battiston et al. (2012) have proposed an alternative measure of systemic risk developed within the complex network framework which they called DebtRank. This new measure takes into account how an institution's (or a group's) financial distress impacts its counterparties across the financial network.

4 Systemic Risk in Complex Financial Networks: Applications

Two distinct empirical methodologies have been developed to use complex networks to analyse issues related to financial stability. One methodology uses simulations for counterfactual analysis and has been widely applied to many types of financial

[18]Extreme conditions within a financial network should be understood as a systemic event, like the recent global financial crisis.

[19]See Castiliognesi and Navarro (2007) for a similar result.

transactions. The other methodology analyses the topological structure of financial networks in order to assess their stability.

4.1 Counterfactual Analysis

One advantage of applying complex network techniques to the analysis of shock propagation is that it allows for counterfactual analysis. For example researchers can simulate the consequences of a failure of a single bank or a group of banks on the rest of the financial system. Upper (2007) provides a detailed description of the methodology.

The starting point of the analysis is the construction of the matrix A, described in Sect. 3, of bilateral exposures representing the graph of the corresponding network. Depending on the level of aggregation the nodes may represent institutions or sectors or countries. In some cases, researchers are able to use data sources that report directly bilateral exposures. However, in many applications researchers have only access to gross measures, for example, the total exposure of each institution to all other institutions. Below we reproduce matrix A but now we have added the level of gross exposures.

$$A = \begin{pmatrix} 0 & 100 & 50 & 0 \\ 50 & 0 & 0 & 100 \\ 0 & 0 & 0 & 50 \\ 100 & 0 & 0 & 0 \end{pmatrix} \begin{matrix} 150 \\ 150 \\ 50 \\ 100 \end{matrix}$$
$$\begin{matrix} 150 & 100 & 50 & 150 \end{matrix}$$

where summing across columns $\sum_j \alpha_{ij}$ gives the total deposits of bank i at all other banks and summing across rows $\sum_i \alpha_{ij}$ gives the total deposits of the banking system in bank j. In order to obtain an estimate of bilateral exposures researchers commonly assume that an entity's assets are spread as evenly as possible given the balance sheet positions of every other entity. Technically, this amounts to the maximization of entropy of the network's linkages. Intuitively, the solution corresponds to the most likely structure of bilateral links given what the researchers know about the level of gross exposures.

Across the various studies that have followed the counterfactual approach there is considerable variation in the level of data aggregation. Below we separate them into two groups, namely, those studies in which nodes represent institutions and those that use higher levels of aggregation.

4.1.1 Risk Exposure at the Institutional Level

The majority of studies that look at the network links of national interbank payments systems use simulations to assess the contagious effects of bank failures. Overall, the evidence of systemic risk is quite mixed. Substantial risk of contagion is reported by Degryse and Nguyen (2007) for Belgium, Van Lelyveld and Liedorp (2006) for the Netherlands, Mistrulli (2011) for Italy, Upper and Worms (2004) for Germany, and Wells (2004) for the UK. In contrast, limited risk of contagion is reported by Blavarg and Nimander (2002) for Sweden, Elsinger et al. (2006) and Lublóy (2005) for Hungary and Sheldon and Mauer (1998) for Switzerland.

Degryse and Nguyen (2007) and Mistrulli (2011) also use their data to assess how structural changes in the banking sector affect systemic risk. The banking systems of both Italy and Belgium have moved away from a relatively complete structure to one dominated by a few banks. The change in network structure seems to have increased the risk of contagion in Italy but to have reduced it in Belgium.

The studies by Angelini et al. (1996) for Italy, Amundsen and Arnt (2005) for Denmark and Furfine (2003) for USA use overnight transactions data and all three report limited possibilities of contagion.

A more general setting is analysed by Anand et al. (2012) who use UK data to calibrate a model that, in addition to links between banks, both domestic and international, also includes links between firms and banks. The authors find that the level of default rates in the corporate sector necessary to trigger a systemic failure in the financial sector is broadly comparable to the rates observed during the Great Depression and those for 2008–2009 during the current global financial crisis.

4.1.2 'Macro' Approaches

Closely related to the structural approach to measuring systemic risk is the work of a strand of the network literature that follows a macro methodology to analyse cross-border financial contagion. The common theme of this work is an abstraction from institutional details by focusing on sector level connections. The level of aggregation that macro approaches use for the construction of networks is significantly higher that used in other applications.

Degryse et al. (2009) use data on bilateral exposures while Castrén and Kavonius (2009) and Castrén and Rancan (2013) calculate bilateral exposures from aggregate exposures by employing the methodology mentioned above. Degryse et al. (2009) analyse the propagation of shocks among the banking sectors of 17 countries on both sides of the Atlantic over the period 1999–2006. In contrast, the other two studies investigate the evolution of the financial interconnectedness of the Euro area. Castrén and Kavonius (2009) use data on financial exposures of seven sectors at the euro area aggregate level over the period 1999–2009 while Castrén and Rancan (2013) use data for the same sectors at the country level.

All three studies strongly suggest that the network structures are time-variant, finding an increase in the number of sector links throughout the period leading to the global financial crisis of 2008–2009. The studies also conclude that factors related to the initial shock, such as geographical location, type of financial instrument, economic sector and country of origin, matter significantly for its potential impact both geographically and quantitatively.

4.2 Topological Structure of Actual Financial Networks

This literature aims to identify and analyse the network topology of the credit links between actual financial institutions. The corresponding theoretical research, reviewed above, clearly suggests that the level of systemic risk crucially depends on the particular structure of the network that includes in addition to topological properties, i.e. random, fully connected, etc., the direction and weights of its links. Empirical implementations of this approach can be very useful given that it allows researchers to identify potential systemic weaknesses in networks that involve thousands of daily transactions between a large number of institutions. However, for a long time, lack of data availability has restricted research to calibrations of theoretical models. This has changed over the last 10 years with the availability of data sets reporting real time bilateral settlements between pairs of institutions in the system. Using data from various types of financial markets located all over the world researchers have identified their corresponding topological structures. Some researchers have then used the structure revealed in this way to perform simulations aiming to identify any risk vulnerabilities within the corresponding payment system.[20]

4.2.1 National Payments Systems

For example, Boss et al. (2004) find that the structure of the Austrian interbank market network (Fig. 2) is characterized by (a) a low clustering coefficient and (b) a short average path length. This is consistent with a highly hierarchical system but as the authors observe this is in contrast to the types of network architectures usually considered in the theoretical literature. Similar conclusions are drawn by Embree and Roberts (2009) for the Canadian payments system, Lublóy (2006) for the Hungarian, Pröpper et al. (2012) for the Dutch system, Sokolov et al. (2012) for the Australian and Soramäki et al. (2007a) for the US Fedwire payments system (Fig. 3). The last of these which examines the degree distribution of the

[20]For a detailed description of the methodology used to perform these simulations see Soramäki et al. (2007a,b). There are some similarities between some of these simulation exercises and the counterfactual methodology reviewed above.

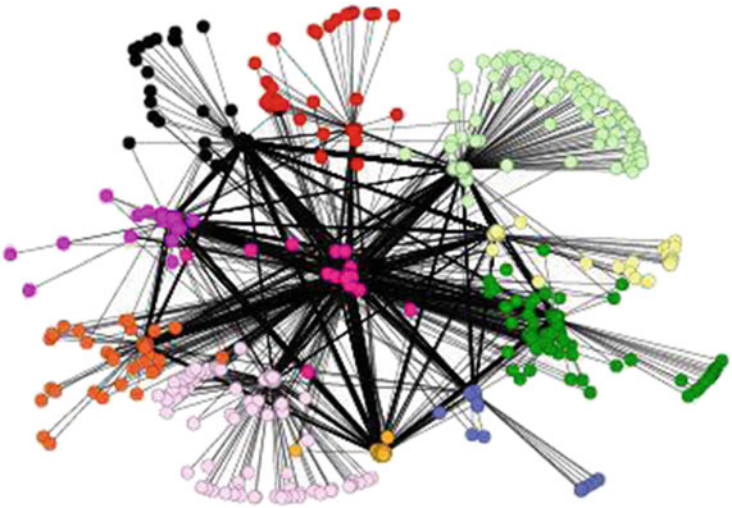


Fig. 2 The Austrian network of interbank payments. *Source*: Boss et al. (2004)

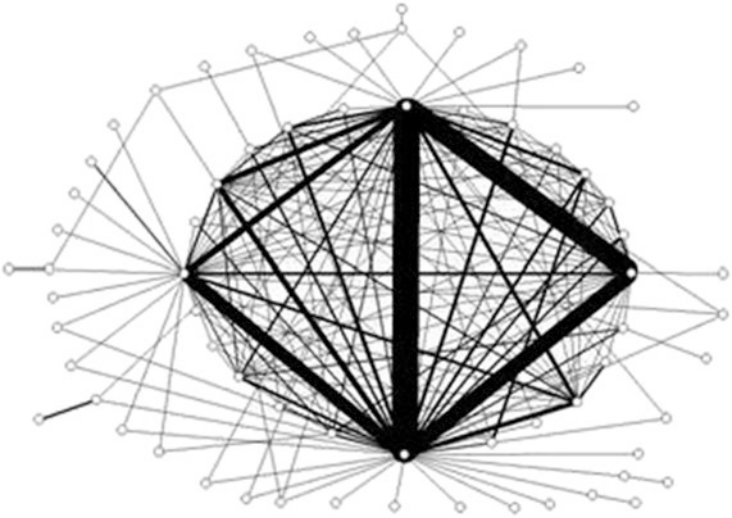


Fig. 3 The US Fedwire payments system. *Source*: Soramäki et al. (2007a,b)

network notices that overall the distribution is scale-free. A fractal structure is also identified by Inaoka et al. (2004) from Japanese banks financial transactions data (Fig. 4). These observations have significant implications for both the likelihood of a systemic failure and the magnitude of its consequences. In particular, it is consistent with the existence of 'too big to fail' institutions that usually are positioned at the top of the hierarchy. The presence of these institutions is responsible for the 'hub

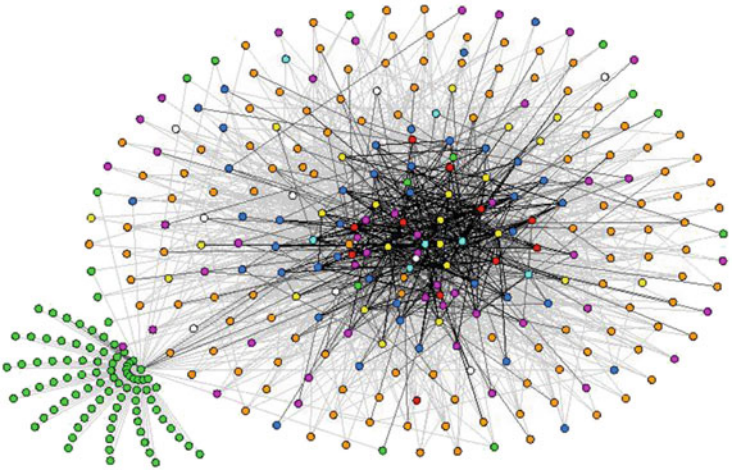


Fig. 4 The Japanese payments system. *Source*: Inaoka et al. (2004)

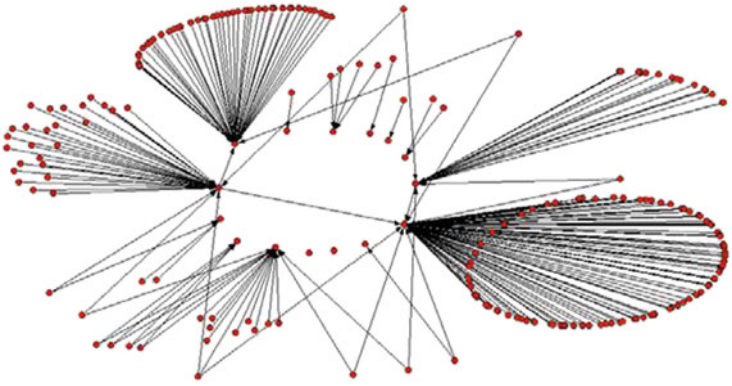


Fig. 5 The UK two-tier system. *Source*: Adams et al. (2010)

and spoke' structure of the graphs of these networks which exhibit short distances between banks even when the frequency of direct relationships is low.

A hierarchical structure is exogenously imposed on the UK system. There is a two-tier system where the top consists of only 15 banks. So-called 'second-tier' banks have to settle their payments through one of the banks in the top tier. Becher et al. (2008) analyse the topological structure of the UK two-tier system (Fig. 5) and compare its stability properties with those of the US system. The authors note that the network structures are similar in the two systems, however, they also find that there are differences in their risk characteristics. In particular, there is a trade-off where the increased risk exposures between the two-tiers is counterbalanced by the benefits derived from (a) liquidity pooling and (b) greater coordination between banks. In a related study Adams et al. (2010) simulate a simple model of network

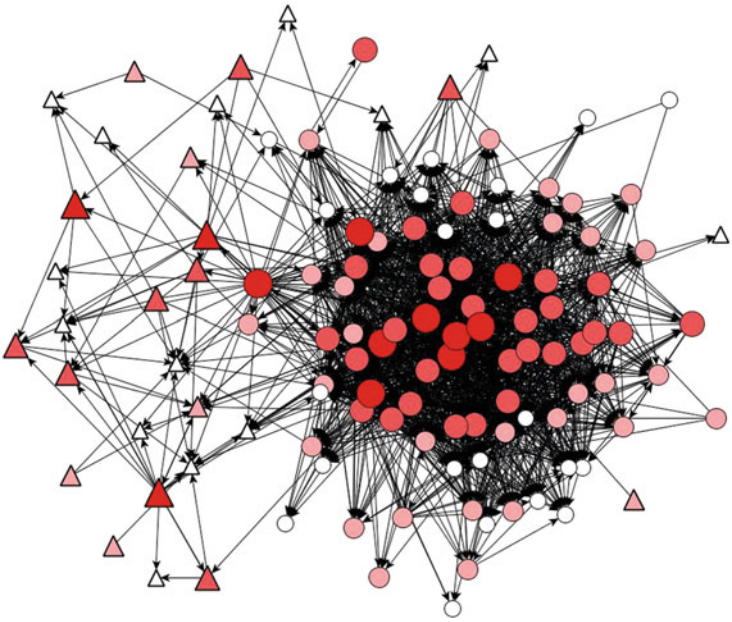


Fig. 6 The Italian payments system. *Source*: Finger et al. (2012)

formation, where each bank can either transact directly with another bank or through a correspondent bank. The authors show that the model is capable of generating a network structure similar to that observed in the UK two-tier system that is driven only by the underlying pattern of payments and the structure of liquidity costs.

The structure of the system can also depend on the level of aggregation of transactions. Finger et al. (2012) using bank transactions from the Italian payments system (Fig. 6) find that when they use daily data the network is random, however, when they move to quarterly frequencies the structure becomes more asymmetric. Put differently, the network exhibits a more hierarchical structure at higher aggregation frequencies, though, still lower than those found for other national systems. They also emphasize that networks constructed at higher aggregation levels have properties that are more stable (less sensitive to the sampling period).

Analyzing the financial stability characteristics of the Colombian payments system, León et al. (2012) offer an example that demonstrates how important it is not to focus exclusively to institutions that are 'too big too fail' but also to those that are 'too interconnected too fail'.[21] Focusing only on the former type of financial institutions in the case of Colombia would have ignored a small-size bank which, because of its high connectivity, imposed potentially a high risk of systemic failure.

[21]For more on this distinction, see Saunders et al. (2009).

4.2.2 Cross-Border Transactions

The study of topological properties is not exclusively restricted to national payments systems. A number of researchers have used the rich transactions data from the Bank of International Settlements to analyse the topology of the network of international payments (Fig. 7). For example, Hattori and Suda (2007) use the dataset to analyse the evolution of the topology of the network of cross-border bank exposures and they find that it has become more tightly connected over time. The network's connectivity, degree and clustering coefficient have increased over time while the average path length has declined. Interestingly, systemic events such as the East Asia twin-crisis in 1997 and the LTCM collapse in 1998 did not disturb this trend. The authors comparing the advantages and disadvantages of these developments suggest that while this higher connectivity offers better opportunities for portfolio diversification and capital allocation it also enhances the likelihood of systemic risk and Haldane (2009) has suggested that over emphasis of the beneficial effects of increased connectivity led economists to under estimate the negative effects of the changes in the other network measures. The evolution of the international payments network is also studied by Garratt et al. (2011). They report that the

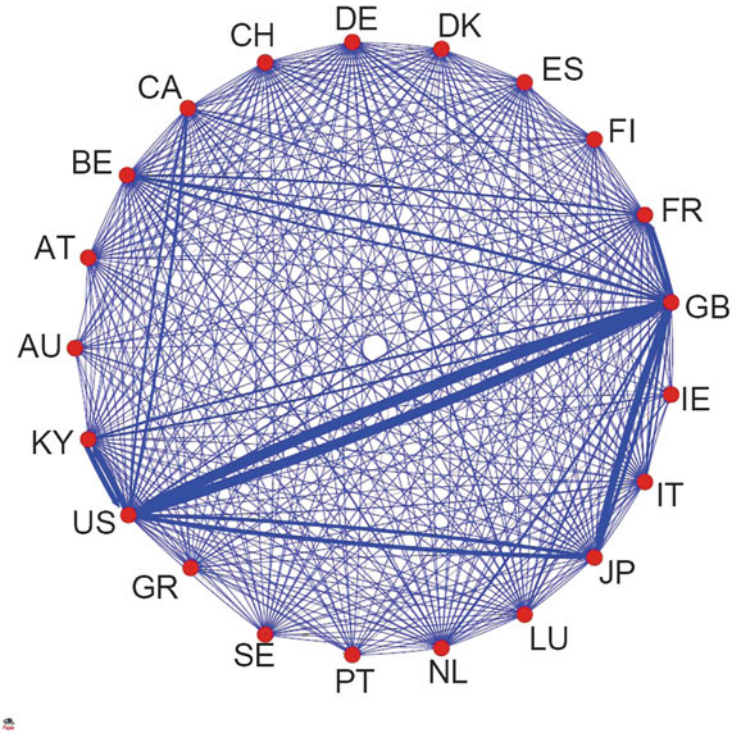


Fig. 7 The International payments system. *Source*: Garratt et al. (2011)

network has changed drastically since the late 1980s. At that time when four financial centres formed one large supercluster the risk of contagion was quite high within its members but much lower globally. Since then the most influential centres have become significantly smaller but the risk of global contagion has significantly grown.

Von Peter (2007) identifies, in the cross-border international payment system, the same hub and spoke structure that characterizes most of the national payment systems. The author also finds that the relatively higher interconnected financial centres, and thus the more important from a systemic risk perspective, are not necessarily the largest in size. McGuire and Tarashev (2008) show how the BIS statistics can be used to identify potentially vulnerable financial centres and provide measurements of their exposure to shocks of various magnitudes.

4.2.3 Other Financial Markets

The majority of research on national financial systems is concentrated on the study of payments systems. Nevertheless, there are still a few applications of the network topology approach to other markets. Rørdam and Bech (2009) compare the networks of the money market (Fig. 8a) and the payments system (Fig. 8b) in Denmark. The payments network consists of banks' proprietary transactions and customer driven transactions while the money market network consists of overnight money market

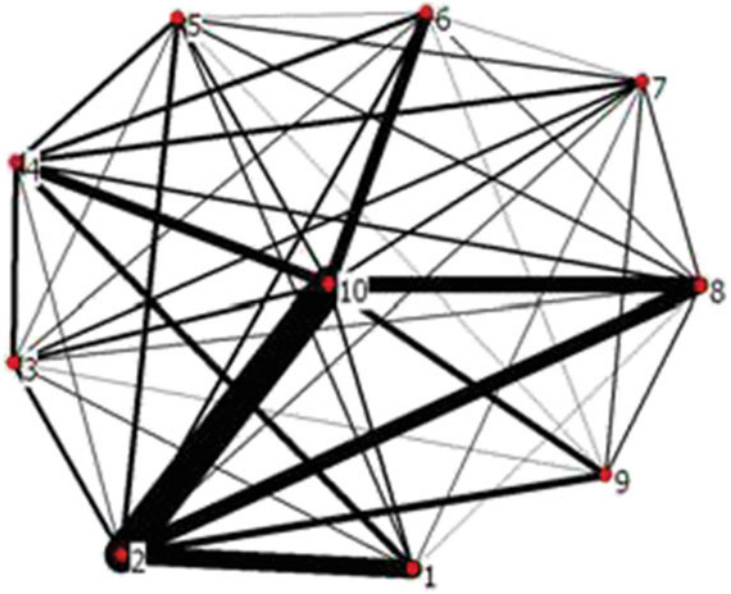


Fig. 8 The Danish banking system. *Source*: Rørdam and Bech (2009)

loans. The authors find that the structures of the two networks differ considerably. While activity in the payments network is dominated by two commercial banks that in the overnight market is more evenly distributed. In contrast to other studies they do not find that the degree distribution of the network is scale-free.

Bech and Atalay (2010) by concentrating on a sub-set of the transactions recorded by the Fedwire system are able to identify the network of the US federal funds market. These are overnight interbank loans that depend on the size of the reserves that the lender holds at the federal reserve system. In accordance with the results reported by Soramäki et al. (2007b) in relation to the total transactions in the Fedwire system, the authors find that the federal funds network exhibits the small-world phenomenon and that the distribution of the number of a bank's counterparties follows a fat-tailed distribution, whereby the vast majority of banks have a few counterparties and a small number of banks have many. However, they also report that the degree distribution of the federal funds market network is not necessarily best represented by a power law distribution.

A similar structure to the one for the federal funds market is reported by Markose et al. (2010) for the US market for Credit Default Swaps (Fig. 9). The market is extremely concentrated with five banks dominating (92 %) the activity in these transactions.

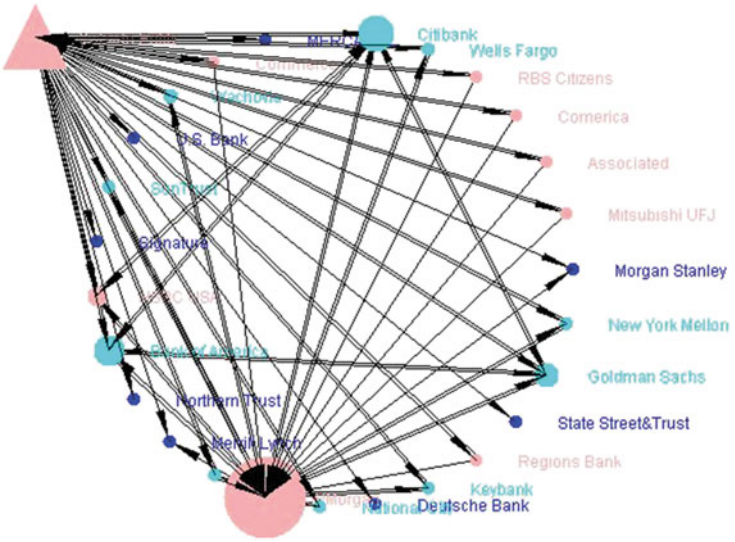


Fig. 9 The US credit default swaps market. *Source*: Markose et al. (2010)

5 Moving Forward

In this paper we have reviewed recent advances in financial economics in relation to the measurement of systemic risk. We have started by reviewing studies that apply traditional measures of risk to financial institutions. Such studies use balance-sheet data to gather information about a financial institution's risk exposure, market data to assess its performance and aggregate data for assigning weights that capture the relative contribution of the institution to systemic risk. Ultimately, the aim is to use reliable measures of contributions to systemic risk for developing pricing schemes that will offer incentives to financial institutions to internalize the externalities that they impose on the rest of the system. Acharya et al. (2009a) describe some interesting proposals aimed at achieving this last goal.[22]

The main focus of the review has been on studies that use network analysis paying special attention to those that apply complex analysis techniques. Applications of these techniques for the analysis and pricing of systemic risk has already provided significant benefits at least at the conceptual level but it also looks very promising from a practical point of view.[23] The basic idea is that what matters for the evaluation of potential systemic consequences of a shock is not only the level of aggregate bilateral risk exposure within the financial system (for example the volume of interbank deposits) but also how this exposure is distributed within the system (the network of interbank deposits). As a first step complex network analysis can be used to trace the contagion effects of a particular shock throughout the system. But perhaps even more importantly it can be used to provide reliable measures for the exposure of the whole system to that shock.

At the theoretical level, significant progress has been made on both of these objectives. Considerable advances have also been made in relation to the first objective at the empirical level. In particular, the topological properties of a wide variety of financial systems around the world, both at the national level and at the level of cross-border transactions, have already been identified. Where improvement is still needed is in using the new tools for deriving practical measures of systemic risk. By practical we mean measures that regulators can use to design insurance pricing schemes for potentially catastrophic events.

The design of optimal schemes would take into account not only the aggregate risk exposure of a network to various kind of shocks but also the incentives that participating institution have to form one type of network over another. The recent work by Acemoglu et al. (2013) does exactly that at the theoretical level. Two issues that have restricted progress at the empirical level are lack of data and measurement problems. We have already discussed above the problem that researchers, who apply

[22]When cross-border risk exposures are significant, as, for example, during the recent global financial crisis, it is also important to ensure there is international coordination among financial regulators; see Acharya et al. (2009b).

[23]Data limitations are a serious constraining factor both for researchers that aim to measure systemic risk and practioners who are interested in controlling it; see Cerutti et al. (2012).

counterfactual analysis, face when they only have data for gross bilateral exposures. At this point, we focus on an alternative challenge that is related to the evaluation of systemic losses as a shock propagates throughout the system.

Clearly, if the losses are restricted to those directly related to the initial shock connectivity would not be a problem. In fact, the more connected a network is, the easier will be to diffuse the losses throughout the system without putting the system itself into danger. What this simplistic analysis ignores is the presence of market imperfections that are responsible for externalities related to (a) fire sales and market freezes (e.g., Acharya et al. 2011; Diamond and Rajan 2011), and (b) liquidity spirals (e.g., Brunnermeier and Pedersen 2009; Garleanu and Pedersen 2007). Earlier, we have reviewed a number of theoretical contributions that allow for these effects within the context of complex network analysis. In contrast, counterfactual and simulations analyses that use actual financial networks as a starting point, usually employ an algorithm developed by Furfine (2003), that assigns an exogenously given parameter (LGD—loss-given-default) each round for these losses. Bridging the gap between theory and applied research in this particular area would add considerably to the attractiveness of the new methodologies.

Finally, some authors have suggested that it might be wise to identify clusters of financial institutions that potentially, because of either/both their size or/and high connectivity, impose disproportional higher risk on the rest of the system. The recent maximum likelihood methods developed by Čopič et al. (2009) for identifying such clusters in network structures might be a promising option.

Acknowledgements We would like to acknowledge financial support from COST Action IS1104 "The EU in the new economic complex geography: models, tools and policy analysis".

References

Acemoglu, D., Ozdaglar, A., & Tahbaz-Salehi, A. (2013). *Systemic risk and stability in financial networks*. NBER Working Paper 18727.

Acharya, V., Gale, D., & Yorulmazer, T. (2011). Rollover risk and market freezes. *Journal of Finance*, *66*, 1175–1207.

Acharya, V., Pedersen, L., Philippon, T., & Richardson, M. (2009a). Regulating systemic risk. In V. Acharya, & M. Richardson (Eds.), *Restoring financial stability* (pp. 283–303). New Jersey: Wiley.

Acharya, V., Pedersen, L., Philippon, T., & Richardson, M. (2010). *Measuring systemic risk*. Unpublished Paper, New York University.

Acharya, V., Wachtel, P., & Walter, I. (2009b). International alignment of financial sector regulation. In V. Acharya & M. Richardson (Eds.), *Restoring financial stability* (pp. 365–376). New Jersey: Wiley.

Adams, M., Galbiati, M., & Giansante, S. (2010). *Liquidity costs and tiering in large-value payment systems*. Bank of England Working Paper 399

Adrian, T., & Brunnermeier, M. (2011). *CoVaR*. NBER Working Paper 17454

Allen, F., & Babus, A. (2009). Networks in finance. In P. Kleindorfer, Y. Wind, & R. Gunther (Eds.), *The network challenge: Strategy, profit, and risk in an interconnected world* (pp.367–382). New Jersey: Pearson Education.

Allen, F., & Babus, A. (2010). *Financial connections and systemic risk*. NBER Working Paper 16177.

Allen, F., & Gale, D. (1998). Optimal financial crises. *Journal of Finance, 53*, 1245–1284.

Allen, F., & Gale, D. (2000). Financial contagion. *Journal of Political Economy, 108*, 1–33.

Amundsen, E., & Arnt, H. (2005). *Contagion risk in the Danish interbank market*. Danmark Nationalbank Working Paper 2005-25

Anand, K., Gai, P., & Marsili, M. (2012). Rollover risk, network structure and systemic financial crises. *Journal of Economic Dynamics and Control, 36*, 1088–1100.

Anand, K., Kirman, A., & Marsili, M. (2011). Epidemics of rules, information aggregation failure and market crashes. *European Journal of Finance*. doi:10.1080/1351847X.2011.601872.

Angelini, P., Maresca, G., & Russo, D. (1996). Systemic risk in the netting system. *Journal of Banking and Finance, 20*, 853–868.

Aubuchon, C., & Wheelock, D. (2010). The geographic distribution and characteristics of U.S. bank failures, 2007–2010: Do bank failures still reflect local economic conditions? *Federal Reserve Bank of St. Louis Review, 92*, 395–415.

Babus, A. (2007). *The formation of financial networks*. Tinbergen Institute Discussion Paper 06-093

Battiston, S., Delli Gatti, D., Gallegati, M., Greenwald, B., & Stiglitz, J. (2007). Credit chains and bankruptcy propagation in production networks. *Journal of Economic Dynamics and Control*, *31*, 2061–2084.

Battiston, S., Puliga, M., Kaushik, R., Tasca, P., & Caldarelli, G. (2012). DebtRank: Too central to fail? Financial networks, the FED and systemic risk. *Scientific Reports, 2*, 541. doi:10.1038/srep00541.

Bech, M., & Atalay, E. (2010). The topology of the federal funds market. *Physica A, 389*, 5223–5246.

Becher, C., Millard, S., & Somaräki, K. (2008). *The network topology of CHAPS Sterling*. Bank of England Working Paper 355

Bhattacharya, S., & Gale, D. (1987). Preference shocks, liquidity and central bank policy. In W. Barnett, & K. Singleton (Eds.), *New approaches to monetary economics* (pp. 69–88). Cambridge: Cambridge University Press.

Billio, M., Getmansky, M., Lo, A., & Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics, 104*, 535–559.

Bitzer, D., & DeMarzo, P. (1992). Sequential banking. *Journal of Political Economy, 100*, 41–61.

Blavarg, M., & Nimander, P. (2002). Inter-bank exposures and systemic risk. *Sveriges Riksbank Economic Review, 2*, 19–45.

Boss, M., Elsinger, H., Summer, M., & Thurner, S. (2004). The network topology of the interbank market. *Quantitative Finance, 4*, 677–684.

Brunnermeier, M., & Oehmke, M. (2012). *Bubbles, financial crises, and systemic risk*. Princeton University Research Paper 47-2012.

Brunnermeier, M., & Pedersen, L. (2009). Market liquidity and funding liquidity. *Review of Financial Studies, 22*, 2201–2238.

Caballero, R., & Simsek, A. (2009). *Complexity and financial panics*. NBER Working Paper 14997.

Castiliognesi, F., & Navarro, N. (2007). *Optimal fragile financial networks*. Tilburg University Discussion Paper 2007-100.

Castrén, O., & Kavonius, I. (2009). *Balance sheet interlinkages and macro-financial risk analysis of the euro area*. ECB Working Paper 1124.

Castrén, O., & Rancan, M. (2013). *Macro-networks: An application to the euro area financial accounts*. ECB Working Paper 1510.

Cerutti, E., Claessens, S., & McGuire, P. (2012). *Systemic risks in global banking: What available data can tell us and what more data are needed?* NBER Working Paper 18531.

Cifuentes, R., Ferrucci, G., & Shin, H. (2005). Liquidity risk and contagion. *Journal of the European Economic Association, 3*, 556–566.

Čopič, J., Jackson, M., & Kirman, A. (2009). Identifying community structures from network data via maximum likelihood methods. *B.E. Journal of Theoretical Economics (BEP), 9*(1), Article 30.

Cossin, D., & Schellhorn, H. (2007). Credit risk in a network economy. *Management Science, 53*, 1604–1617.

Da Cruz, J., & Lind, P. (2012). The dynamics of financial stability in complex networks. *European Physical Journal B, 85*, 256.

De Jonghe, O. (2010). Back to the basics in banking? A micro-analysis of banking system stability. *Journal of Financial Intermediation, 19*, 387–417.

Degryse, H., & Nguyen, G. (2007). Interbank exposures: An empirical examination of systemic risk in the Belgian banking system. *Journal of International Central Banking, 3*, 123–171.

Degryse, H., Elahi, M., & Penas, M. (2009). *Cross-border exposures and financial contagion*. European Banking Center Discussion Paper 2009-02.

Diamond, D., & Dybvig, P. (1983). Bank runs, deposit insurance and liquidity. *Journal of Political Economy, 91*, 401–419.

Diamond, D., & Rajan, R. (2011). Fear of fire sales, illiquidity seeking, and credit freezes. *Quarterly Journal of Economics, 126*, 557–591.

Eisenberg, L., & Noe, T. (2001). Systemic risk in financial systems. *Management Science, 47*, 236–249.

Elliott, M., Golub, B., & Jackson, M. (2013). *Financial networks and contagion*. Available at SSRN http://ssrn.com/abstract=2175056.

Elsinger, H., Lehar, A., & Summer, M. (2006). Risk assessments for banking systems. *Management Science, 52*, 1301–1314.

Embree, L., & Roberts, T. (2009). *Network analysis and Canada's large value transfer system*. Bank of Canada Discussion Paper 2009-13.

Erdös, P., & Rényi, A. (1959). On random graphs. *Publicationes Mathematicae Debrecen, 6*, 220–297.

European Central Bank. (2007). In *Conference Proceedings of the Fourth Joint Central Bank Research Conference*, November 8–9, 2005, ECB, Frankfurt.

Finger, K., Fricke, D., & Lux, T. (2012). *Network analysis of the e-MID overnight money market: The informational value of different aggregation levels for intrinsic dynamic processes*. Kiel Working Paper 1782.

Furfine, C. (2003). Interbank exposures: Quantifying the risk of contagion. *Journal of Money, Credit and Banking, 35*, 111–128.

Gai, P., & Kapadia, S. (2010). Contagion in financial networks. *Proceedings of the Royal Society A, 466*, 2401–2423.

Garleanu, N., & Pedersen, L. (2007). Liquidity and risk management. *American Economic Review Papers and Proceedings, 97*, 193–197.

Garratt, R., Mahadeva, L., & Svirydzenka, S. (2011). *Mapping systemic risk in the international banking network*. Bank of England Working Paper 413.

Gauthier, C., Lehar, A., & Souissi, M. (2009). *Macroprudential capital requirements and systemic risk*. Unpublished Paper, Bank of Canada.

Giglio, S. (2011). *Credit default swap spreads and systemic financial risk*. Unpublished Paper, Harvard University.

Goyal, S. (2009). *Connections: An introduction to the economics of networks*. Princeton: Princeton University Press.

Gray, D., Merton, R., & Bodie, Z. (2008). *New framework for measuring and managing macrofinancial risk and financial stability*. Harvard Business School Working Paper 09-015.

Haldane, A. (2009). *Rethinking the financial network*. Speech delivered at the Financial Student Association, Amsterdam.

Haldane, A., & May, R. (2011). Systemic risk in banking ecosystems. *Nature, 469*, 351–355.

Hartmann, P., Straetmans, S., & de Vries, C. (2005). *Banking system stability: A cross-Atlantic perspective*. NBER Working Paper 11698.

Hattori, M., & Suda, Y. (2007). *Developments in a cross-border bank exposure "network"*. Bank of Japan Working Paper 07-E-21.

Huang, X., Zhou, H., & Zhu, H. (2012). Systemic risk contributions. *Journal of Financial Services Research, 42*, 55–83.

Inaoka, H., Ninomiya, T., Taniguchi, K., Shimizu, T., & Takayasu, H. (2004). *Fractal Network derived from banking transactions: An analysis of network structures formed by financial institutions*. Bank of Japan Working Paper 04-E-04.

Iori, G., Jafarey, S., & Padilla, F. (2006). Systemic risk on the interbank market. *Journal of Economic Behavior and Organization, 61*, 525–542.

Jackson, M. (2008). *Social and economic networks*. Princeton: Princeton University Press.

Kiyotaki, N., & Moore, J. (2004). *Credit chains*. University of Edinburgh ESE Discussion Paper 118.

Lagunoff, R., & Schreft, L. (2001). A model of financial fragility. *Journal of Economic Theory, 99*, 220–264.

Lehar, A. (2005). Measuring systemic risk: A risk management approach. *Journal of Banking and Finance, 29*, 853–864

Leitner, Y. (2005). Financial networks: Contagion, commitment, and private sector bailouts. *Journal of Finance, 60*, 2925–2953.

Lenzu, S., & Tedeschi, G. (2012). Systemic risk on different interbank network topologies. *Physica A, 391*, 4331–4341.

León, C., Machado, C., Cepeda, F., & Sarmiento, M. (2012). Systemic risk in large value payment systems in Colombia: A network topology and payments simulation approach. In M. Hellqvist, & T. Laine (Eds.), *Diagnostics for the financial markets: Computational studies of payment system* (pp. 267–313). Helsinki: Edita Prima Oy.

Lublóy, A. (2005). Domino effect in the Hungarian interbank market. *Kozgazdasagi Szemle (Economic Review), 42*, 377–401.

Lublóy, A. (2006). *Topology of the Hungarian large-value transfer system*. Magyar Nemzeti Bank Occasional Paper 57.

Markellof, R., Warner, G., & Wollin, E. (2012). *Modeling systemic risk to the financial system*. MITRE Corporation Technical Paper 12-1870.

Markose, S., Giansante, S., Gatkowski, M., & Shaghaghi, A. (2010). *Too interconnected to fail: Financial contagion and systemic risk in network model of CDS and other credit enhancement obligations of US banks*. University of Essex Discussion Paper 683.

May, R., & Arinaminpathy, N. (2010). Systemic risk: The dynamics of model banking systems. *Interface: Journal of the Royal Society, 7*, 823–838.

May, R., Levin, S., & Sugihara, G. (2008). Ecology for bankers. *Nature, 451*, 893–895.

McGuire, P., & Tarashev, N. (2008). *Global monitoring with the BIS international banking statistics*. BIS Working Paper 244.

Mistrulli, P. (2011). Assessing financial contagion in the interbank market: maximum entropy versus observed interbank lending patterns. *Journal of Banking and Finance, 35*, 1114–1127.

Montagna, M., & Lux, T. (2013). *Hubs and resilience: Towards more realistic models of the interbank markets*. Kiel Working Paper 1826.

Nier, E., Yang, J., Yorulmazer, T., & Alentorn, A. (2007). Network models and financial stability. *Journal of Economic Dynamics and Control , 31*, 2033–2060.

Pröpper, M., van Lelyveld, I., & Heijmans, R. (2012). Network dynamics of TOP payments. In M. Hellqvist & T. Laine (Eds.), *Diagnostics for the financial markets: Computational studies of payment system* (pp. 235–266). Helsinki: Edita Prima Oy.

Rørdam, K., & Bech, M. (2009). The topology of Danish interbank money flows. *Banks and Bank Systems, 4*, 48–65.

Rochet, J-C., & Tirole, J. (1996). Interbank lending and systemic risk. *Journal of Money, Credit, and Banking, 28*, 733–762.

Saunders, A., Smith, R., & Walter, I. (2009). Enhanced regulation of large complex financial institutions. In V. Acharya & M. Richardson (Eds.), *Restoring financial stability* (pp. 139–156). New Jersey: Wiley.

Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A., & White, D. (2009). Economic networks: The new challenges. *Science, 325*, 422–425.

Segoviano, M., & Goodhart, C. (2009). *Banking stability measures*. IMF Discussion Paper 627.

Sheldon, G., & Mauer, M. (1998). Interbank lending and systemic risk: An empirical analysis of Switzerland. *Swiss Journal of Economics and Statistics, 134*, 685–704.

Sokolov, A., Webster, R., Melatos, A., & Kieu, T. (2012). Loan and nonloan flows in the Australian interbank network. *Physica A, 391*, 2867–2882.

Soramäki, K., Bech, M., Arnold, J., Glass, R., & Beylerer, W. (2007a). The topology of interbank payment flows. *Physica A, 379*, 317–333.

Soramäki, K., Beyeler, W., Bech, M., & Glass, R. (2007b). New approaches for payment system simulation research. In H. Leinonen (Ed.), *Simulation studies of liquidity needs, risks and efficiency in payment networks* (pp.15–39). Helsinki: Edita Prima Oy.

Tarashev, N., Borio, C., & Tsatsaronis, K. (2010). *Allocating systemic risk to individual institutions; methodology and policy applications*. BIS Working Paper 308.

Teteryatnikova, M. (2012). *Systemic risk in banking networks: Advantages of "Tiered" banking systems*. University of Vienna Working Paper 1203.

Upper, C. (2007). *Using counterfactual simulations to assess the danger of contagion in interbank markets*. BIS Working Paper 234.

Upper, C., & Worms, A. (2004). Estimating bilateral exposures in the German interbank market: Is there a danger of contagion? *European Economic Review, 48*, 827–849.

Van Lelyveld, I., & Liedorp, F. (2006). Interbank contagion in the Dutch banking sector: A sensitivity analysis. *International Journal of Central Banking, 2*, 99–134.

Varela, L-M., Rotundo, G., Ausloos, M., & Carrete, J. (2015). Complex networks analysis in socioeconomic models. In P. Commendatore, S. Kayam, & I. Kubin (Eds.), *Complexity and geographical economics: Topics and tools*. Heidelberg: Springer (This volume).

Vega-Redondo, F. (2007). *Complex social networks*. Cambridge: Cambridge University Press.

Von Peter, G. (2007). *International banking centres: A network perspective* . BIS Quarterly Review Working Paper, http://dx.doi.org/10.2139/ssrn.1075205.

Wells, S. (2004). *Financial interlinkages in the United Kingdom's interbank market and the risk of contagion*. Bank of England Working Paper 230.

Wong, A., & Fong, T. (2010). *An analysis of the interconnectivity among the Asia-Pacific economies*. Unpublished Paper, Hong Kong Monetary Authority.

Migration and Networks

Douglas R. Nelson

Abstract This paper provides a brief overview of current research on networks in international migration. The paper begins with a short discussion of the relationship between networks and social capital. While controversial, this concept potentially provides a unifying thread linking both various aspects of economic research and, potentially more importantly, providing a bridge linking economic research to parallel research in demography and sociology. The core of the paper is a discussion of the role of networks in the decision to migrate, the role of networks in assimilation, and the effect of global migrant networks on the pattern of international trade. In all three of these areas, recent years have seen substantial new research, both theoretical and empirical, on the ways networks interact with more standard economic variables. In each of these cases, networks are seen to play an essential role in the migration experience.

1 Introduction

Migration is an economic and, more broadly, social phenomenon of profound importance. In quantitative terms, international migrants are about 3 % of current global population (about 214 million people).[1] Even as a proportion of the world's population, this is in the neighborhood of the largest migrations in the historical record. Even if we only focus only on wage differentials, the impact of migration on the migrants themselves is huge. The impact on the sending and receiving communities are matters of considerable dispute, but it seems reasonable to conclude that barriers to arbitrage of human capital between national labor markets

[1]This is the UN, Department of Economic and Social Affairs, Population Division, estimate of 2010 migration stock from their 2008 *Trends in International Migration Stock: The 2008 Revision*, Table 1.

D.R. Nelson (✉)
Murphy Institute, Tulane University, 108 Tilton Hall, New Orleans, LA 70118-5698, USA
e-mail: dnelson@tulane.edu

P. Commendatore et al. (eds.), *Complexity and Geographical Economics*,
Dynamic Modeling and Econometrics in Economics and Finance 19,
DOI 10.1007/978-3-319-12805-4_7

are considerably larger than those facing final commodities or capital (Hamilton and Whalley 1984; Walmsley et al. 2011).

The fact that international migration involves the movement of people (i.e. members of families and communities with distinctive cultural, social and political attributes along with their economic differences) means that migration has first-order implications extending far beyond those of international trade in commodities and international capital flows. One way of beginning to broaden our perspective on migration so as to incorporate these broader social considerations within our economic analyses is to draw on network analytic methods and results. Sociologists and demographers have long taken networks to be an essential aspect of the migration decision and sought to systematically incorporate them in their research. This is now true of research by economists as well. This chapter provides a brief overview of research on migration that explicitly focuses on networks as they bear on: the decision to migrate; the selection of host country; assimilation (in particular, labor market success); and the link between migration and international trade. Before turning to these issues, we begin with a short discussion of some concepts that will be deployed throughout this chapter.

2 Networks, Social Capital and Migration

One of the great virtues of a network analytic approach to the analysis of social interaction is that it permits both a focus on individual nodes (i.e. agents) and on the overall social context as reflected in the structure of the network in which the individual nodes are embedded. For example, it is useful to know how many links (i.e. relationships) an individual has (i.e. the node's "degree"), or whether those links involve other agents who are well-connected, or poorly connected, or how well connected a node is (i.e. how long are the shortest paths from a given node to any other node); but it is also interesting to know about aggregate properties like the degree distribution, the average path length between nodes, or the shortest or longest shortest path across all nodes.[2]

Even when individual agents are the object of study, network analysis is fundamentally about the web of social relations in which those agents are embedded. The issue is always the links. That is, understanding the content and meaning of the links is a central part of any network analysis. From a descriptive point of view, these links are often flows of various resources (information, money, etc.). Analytically, though, the links are intended to represent social relations of some

[2]Graph theory is the language of network theory. Unfortunately, there is no really standard language of graph theory. Nonetheless, there are some pretty standard usages. For example, the shortest path between two nodes is usually called the "distance" or "geodesic". The "eccentricity" is the largest distance for a given node. The "radius" is the smallest eccentricity among all nodes, while the "diameter" is the largest. Chapter 1 of Harris et al. (2008) is an excellent introduction to the basics of graph theory.

kind (kinship, friendship, authority, power, etc.), and the global structure of those links is seen as an object of study on its own. Thus, for example, one might analyze international migration flows using network analytic tools. That is, one could calculate the standard measures of network topology: degree distribution (from which one can test whether the network is scale free); average path length; centrality; and clustering; among others.[3] The patterns identified by this sort of analysis can then serve as an input to further analyses (as either dependent or independent variables).

Probably the most common theoretical framework for linking the analysis of networks to migration is social capital theory. Broadly speaking, social capital refers to the resources available to individuals by virtue of their membership in a well-defined group.[4] Especially in environments where resources cannot be effectively distributed by arms-length methods (e.g. markets), relatively dense social networks that rely on trust and community enforcement can support collectively and individually better outcomes than would be available in the absence of such relations. As argued by Burt (2005), location in a network is likely to be as important to an individual as simple membership. Specifically, Burt develops an analysis in terms of "brokerage" and "closure". The latter refers to the presence of a relatively dense network of strong connections that support trust, reciprocity and alignment of interests. By contrast, the former refers to a specific type of structural position in a network—one that spans "structural holes" (Burt 1992, 2002). That is, the individual is able to mediate relations between members of the group that are not intensively linked with one another. Much of Burt's work attempts to demonstrate the particular value of brokerage in certain types of organizational structure (i.e. business firms). That is, human capital differs for individual members of the same network depending on their position in that network. However, Burt (2001, 2009) also argues that closure plays an essential role in the operation of networks. In particular, closure underwrites trust among members in the network. Similar considerations play a role in the macro analysis of social capital. For example, as we shall see, the presence of a relatively dense network spanning more than one market is able to replace missing markets. The key in such a case is that the agents carrying out the brokerage must be sufficiently embedded in a network characterized by a sufficiently high degree of closure that it supports transactions requiring a high degree of trust.

[3] Interestingly, while there are many papers examining international trade flows in this fashion, and global migration has long been analyzed as interacting systems (Kritz et al. 1992), I know of no studies of the network topology of global migration.

[4] There are a number of problems with social capital as a concept, perhaps more importantly, it is not clear that its creation has the instrumental properties of the other forms of capital (Solow 2000). That is, most of the groups considered in research on social capital (e.g. ethnic communities) are not created for the purpose of generating social capital. Analytically, this makes social capital an endowment, like capital in static economic analysis. On the other hand, social capital is produced and reproduced via the actions of members of the group (even though many of these actions may not be instrumental from the point of view of the members in any obvious way).

Unlike physical, financial and human capital, social capital is inherently a property/product of groups, rendering straightforward identification difficult. In this regard, the link to networks has proved useful. Lin (2001) even defines social capital as "resources embedded in social networks accessed and used by actors for actions". This raises the possibility that we can use the observable network properties of a group of people as a way of getting at social capital. Unfortunately, even this seems tricky—which properties are to be associated with social capital? Part of the problem has to do with whether we want to view social capital as fundamentally an attribute of individuals or of groups. From the point of view of individuals, we have already noted, following Burt, that structural heterogeneity is essential. However, from the point of view of the group, it seems clear that aggregate properties will be the main focus of analysis.

3 Networks and the Migration Decision

The decision to migrate, including the selection of destination, is probably the most studied topic in research on migration. Much of the early work involved listing of "pull" and "push" factors as a prelude to essentially *ad hoc* analysis of individual or aggregate data. For economists, the usual approach was, and is, to see migration as labor arbitrage (e.g. Hicks 1932). This basic idea was substantially extended in Sjaastad's (1962) now classic application of human capital theory to the migration decision. The idea is to see migration as an investment decision—i.e. an agent considers two "assets": the net present value of labor earnings net of costs of staying in the source country; and the net present values of labor earnings net of costs of working in a foreign country. Sjaastad was primarily interested in examining the various elements of costs and benefits that enter the calculation and, thus, presented the analysis primarily in terms of a single agent and the implications for aggregate (especially gross) migration flows. This logic has been developed in a number of directions.

For our purposes, the most immediately relevant development was a shift in perspective from the individual migrant to the family as the basic decision-making unit. Mincer (1978) took a new household economics perspective and analyzed the decision of whether or not to move the entire family. Stark and Levhari (1982) recognize that by treating the decision-making unit as the family (instead of the individual), allocation of family members between markets whose economies face imperfectly correlated shocks (or business cycles) can be seen as a more-or-less standard portfolio allocation problem. Empirical work on remittance flows is strongly consistent with this logic (e.g. Yang and Choi 2007; Yang 2008). As we shall see, the family plays a major role in the network analytic approach to migration.

Networks primarily enter the economic analysis of the migration decision on the cost side of the household's calculation.[5] The two most prominent ways that networks affect costs are through provision of information and provision of financial resources. With respect to the latter, there is substantial research showing the large magnitude of such flows in the form of remittances (Yang 2011).[6] Those resources can be used in myriad ways, but one potential use is funding of additional migration by family (or more broadly community) members. Similarly, reliable information about migration channels, labor market conditions, housing, and the like, can reduce uncertainty dramatically. While the finance channel is surely important, most research on networks and migration emphasize the information channel (e.g. Teteryatnikova 2013). Not surprisingly, historical narratives stress both of these as sources of what has long been called "chain migration" (Jones 1992; Wegge 1998). While the provision of information and finance has long been understood as a fundamental part of migration systems, only recently have researchers begun to systematically study these in the context of models linked to an explicit notion of networks. In recent years, a number of theoretical papers have sought to model the link between networks (usually defined in terms of flows of family or community members), the cost of migration and migration dynamics. While focused on interregional migration, Carrington et al. (1996) is an important, early example of this literature.[7] In that paper, the authors build a (discrete time) dynamic model of the sort implied by Sjaastad, with the additional feature that the cost of migration is decreasing in the number of migrants already in the host country. As in all of the later papers of this type, one of the keys to understanding the model is the recognition that the cost reduction due to additional migration is an externality. Not surprisingly, this model permits a variety of steady states, including: complete migration; positive, but less than complete, migration; multiple equilibria involving both low and high migration. More importantly for interpreting empirical results, this model also involves continuing migration even as wages equalize.

Even before systematic models of migration decision-making with network effects were developed, empirical work on migration choice had identified networks as a key factor in the migration decision. Empirical work on migration falls, loosely speaking, into two groups based on the nature of the data used: microeconomic

[5]In a broader sense, of course, networks can easily be seen to work on the benefit side as well. Most obviously, people might choose to migrate to a place where there are family or community members because they want to be near those particular people.

[6]Stark and Jakubek (2013) present a theoretical analysis of the role of migration networks in the financial support of migration and analyze the implications for optimal network size. Interestingly, there is survey based evidence that the more densely embedded in a network of co-ethnics is an agent, the more likely they are to remit earnings (Chort et al. 2012).

[7]Chau (1997) presents a similar analysis. Teteryatnikova (2013) builds on Calvo-Armengol and Jackson's (2004) model of job search to provide explicit network microfoundations for models of the Carrington, et al. sort in terms of information provision. Spilimbergo and Ubeda (2004) present an analysis emphasizing family relations, but their causal structure runs through the utility function (a preference to be with "friends and family") rather than through network effects.

data derived from household surveys; and macroeconomic data based (usually) on government collected data on aggregate stocks and flows of migrants. With respect to microeconomic data, the most common source of data, and inspiration, is Jorge Durand and Douglas Massey's Mexican Migration Project (MMP). Since the early 1980s, this project has carried out household surveys in Mexican villages (Massey et al. 1987).[8] From the beginning this project has studied the effect of networks on migration decisions (Taylor 1986; Massey 1990; Massey and Espinosa 1997; Zahniser 1999; Phillips and Massey 2000; Deléchat 2001; Palloni et al. 2001; Durand and Massey 2004; Fussell and Massey 2004; Fussell 2004; Orrenius and Zavodny 2005; Bauer et al. 2007; Dolfin and Genicot 2010; Flores-Yeffal and Aysa-Lastra 2011; Massey et al. 2011; Flores-Yeffal 2013). This work consistently finds that network effects are a statistically and economically significant factor in both the decision to migrate and the selection of a migration location. Similar results are found using survey data from alternative Mexican sources (Winters et al. 2001; Davis et al. 2002; McKenzie and Rapoport 2007, 2010), Ecuador (Bertoli 2010), Germany (Bauer and Zimmermann 1997; Haug 2008; Kanas et al. 2012), Denmark (Nannestad et al. 2008), and Hong Kong (Wong and Salaff 1998). In addition, similar results have been found for intra-country studies in, among others the US (Choldin 1973), India (Banerjee 1983), China (Zhao 2003) and Thailand (Garip 2008). This work provides strong support for theoretical models of the sort discussed in the previous paragraph. In particular, it is clear that early migrants appear to lower the cost of later migrants, leading to chain migration (also often called cumulative causation of migration). Interestingly, it also appears that, as the level of migration from a given community rises, community networks substitute for family networks (that is, past some threshold, family links add no additional explanatory power to the presence of community links).

Where survey based research can ask very detailed questions about the nature of network (e.g. family, community, etc.) and about individual decision context (e.g. family size, assets, relative wealth, etc.), it is, in its nature, very specific. Research based on aggregate data lacks this detail, but permits much broader application in terms of source and host countries. There is a sizable literature that simply introduces a migration stock variable into a gravity (or gravity-like) regression (Levy and Wadycki 1973; Bao et al. 2009; Jayet et al. 2010; Marques 2010). This, essentially *ad hoc* work, consistently finds that networks (proxied by stocks of migrants from the sending country) are a significant explanatory variable.[9] The

[8] According to their webpage, the MMP "comprises 143 communities with 22,894 households surveyed in Mexico and 957 households surveyed in the United States. It provides individual level data on 75,066 males and 76,714 females, for a total of 151,785 persons". More recently this methodology has been extended to a number of Caribbean, Central and South American countries under the Latin American Migration Project (LAMP).

[9] The gravity model has long been the workhorse empirical framework for empirical work on migration using aggregate data. As with networks, it is quite standard to simply include a variable *ad hoc* to evaluate its significance in explaining migration flows (e.g. welfare generosity, more or less strict immigration laws, etc.). The same is also true in the case of empirical trade research. The

big advance in aggregate empirical work on migration came with Borjas' (1987) paper on agent heterogeneity and the implications of selection for research and policy. Rather than the broad implication that net migration between two countries should be positive if the wage differential (net present value of income differentials) favors the source country, Borjas shows that, in addition to the difference in mean wages (adjusted for costs of migration), the variance of wages (relative inequality) in each country matters. Borjas' paper (see also Tunali 2000, written about the same time) built directly on Heckman's (1979) fundamental analysis of the selection problem, making clear the implication for aggregate empirical work (as well as the appropriate econometric response).[10] Pedersen et al. (2008) incorporate both selection and network effects in a straightforward gravity model, finding that network effect dominate selection effects in that setup. McKenzie and Rapoport (2010) develop an explicit analysis of network effects in the context of the Roy-Borjas model. Their key result is that, for small networks migrants are somewhat positively selected, but as networks grow (reducing costs), relatively poor migrants are increasingly able to afford migration, resulting in negative selection. Overall, network effects seem to play a very significant role in determining whether or not to migrate, and seem to play a dominant role in determining migration target.

4 Neighborhoods, Networks and Assimilation

From a broad social point of view, assimilation is perhaps the single most important issue in thinking about migration. Loosely speaking, "assimilation" refers to the process of immigrants becoming more like natives (or the state of being more like natives). However, operationally it is a nearly hopelessly broad concept, including language acquisition, adoption of broad cultural norms, similar educational attainment to natives, and the like. All of these have been studied at length (mostly by sociologists). When economists talk about assimilation, we mostly refer to labor market norms: wages, unemployment, etc. Whatever the form of assimilation, research on this topic is highly contested.[11] This is not the place to review the broad literature on assimilation. Instead, we briefly discuss research on the role of networks in supporting/resisting assimilation.

---

point is not that gravity models themselves are *ad hoc*, but, rather, that whatever foundations one is using, the empirical performance is so good that one is tempted to include a variety of *ad hoc* variables that seem potentially important.

[10]It is also well worth looking at Heckman and Honore's (1990) analysis of the Roy model in this context.

[11]Even the question of whether assimilation is, or is not, a "good thing" is contested. Since much of this literature is driven by normative concerns, it is not surprising that even the facts of the matter are highly contested.

We have just seen that networks play a major role in getting people from one place to another. Different, but surely overlapping, networks are equally important in helping migrants find housing, jobs and, difficult to measure but surely of considerable importance, comfort in a foreign (and often hostile) land. In network theoretic terms, broad notions of social capital imply that there is sufficient density of linkage within a network for common information and expectations to support extended reciprocity. However, as sociologists emphasize, common culture also eases entry of new members into a local network. This is partly due to easier communication, common normative expectations, and more portable reputations. Thus, newly arrived migrants need not be known widely in the network to reap the benefits of participation in the network (Portes and Stepick 1993; Portes and Rumbaut 2006).

The key issue in research on networks/social capital in the assimilation of migrants is whether, loosely speaking, they act as springboards or traps. On the one hand, by offering access to financial, informational, and social support that would not be available on the same terms to immigrants, immigrant networks must have a positive effect on both individual members of the network and, thus, on the overall performance of the community. Especially work that views networks through the lens of social capital tends to take the "springboard" view (e.g. Coleman 1988; Lin 2001; Burt 2005). In this work, social capital is about access to resources that are available to members of a given network. Ethnographic studies of immigrant networks tend to emphasize these benefits.[12]

The other side of the coin, however, is that these resources are available because of the social relations among members of the network, and it is here that closure plays a particularly significant role. By protecting collective resources from opportunistic exploitation, closure ensures continuing access of those resources to members of the community. Enforcement in this context includes exclusion from both tangible resources that are the usual focus of network/social capital theories and the more general emotional support deriving from membership in a community. As a practical matter, what this means is that members of the network need to engage in behaviors that are reproductive of the network. This may involve general social behaviors like attending particular churches or community celebrations, but may also involve behaviors inconsistent with market logics. This, of course, is precisely the point: a broader social logic displaces pure market logic. If this results in hiring less skilled workers, or more workers (from the community) than the cost-minimizing number, the result could be lower productivity and lower profits than in firms outside the network. Similarly, if workers need to maintain linguistic and other cultural markers that identify them as community members, but also identify them as outsiders to native employers, the result could easily be poorer

[12]Portes and Sensenbrenner (1993) provide a useful review of the relevant sociological literature. It should be noted, however, that Portes and Sensenbrenner also discuss the negative effects of social embeddedness at length.

labor market performance (Chiswick 1978; Dustmann 1994; Chiswick and Miller 1996; Dustmann and Soest 2001).

There are many challenges facing empirical research on these issues—e.g. identifying meaningful networks/communities, measuring density of links (or any other network topological properties), measuring performance, forging a clean identification of the relationship between degree of social capital and performance. Current research does seem to suggest that the larger the ethnic community ("enclave") the more slowly to members of the community develop local language skill and, presumably, other host country behavioral markers (Chiswick 1991; Cutler and Glaeser 1997; Borjas 1998; Lazear 1999; Chiswick and Miller 2005; Cutler et al. 2008).[13] This leads some to conclude that immigrant enclaves result in poor assimilation (Borjas 1994, 1995, 2000; Lazear 2007; Warman 2007; Xie and Gough 2011; Danzer and Yaman 2013). There are, however, difficult endogeneity issues here. For example, if the biggest gains from participation in networks go to relatively unskilled co-ethnics, we would expect to see enclaves characterized by relatively low levels of skill, education, language acquisition, and the like. Studies that have focused on these issues tend to find more positive effects from participation in the enclave economy for unskilled workers (Wilson and Portes 1980; Edin et al. 2003; Mahuteau and Junankar 2008; Damm 2009). Similarly mixed results are found when one looks at the effect of enclaves on the development of immigrant entrepreneurs (Sanders and Nee 1987; Portes and Jensen 1989; Sanders and Nee 1992; Light et al. 1994; Portes and Zhou 1996, 1999; Aguilera 2009).

More than in the case of the migration decision, when we seek to understand the relationship between network and assimilation we are faced with the importance of understanding the meaning of the links and not simply their topology. The study of capitalist markets is greatly eased by the extent to which real markets approximate the ideal of relatively asocial relation in those markets.[14] Once we admit a more broadly socialized domain of economic decision-making, we need to be much more aware of shared meanings and specific (as opposed to abstract) social relations.

[13] An exception is recent work by Munshi (2003, 2011), who develops models of migration with job search and occupational traps, respectively, in which networks increase the likelihood of positive outcomes. In both cases, the empirical work presented by Munshi is consistent with the implications of the models. A very useful review of the literature on networks, neighborhoods and job search that covers many of the topics, but without a focus on migration is Ioannides and Loury (2004).

[14] Of course, real markets are characterized by varying mixes of social content (Kirman 2011). It remains the case that, as a wide variety of analysts have suggested, the disembedding of the market is one of the truly distinctive features of capitalism and the extent to which that disembedding fails is a potential source of crisis in the overall capitalist system (e.g. Schumpeter 1942/1975; Polanyi 1994/2001; Habermas 1975).

5 Migration Networks and International Trade

To this point, we have focused on first-order relations between networks and migration. Another way of examining the link between networks and migration is to examine how migrant networks affect some other economically relevant phenomenon. There is a surprisingly large literature on the link between migration and trade, and one branch of this literature has seen the relationship as being about networks from the earliest papers on this topic. Specifically, empirical work on this question tends to view migration as a factor potentially reducing the cost of trade between countries. Not surprisingly, the gravity model as applied to trade has been seen as a natural econometric framework for such empirical work (for surveys, see Anderson and VanWincoop 2004; Anderson 2011; Bergstrand and Egger 2011). It is well known that the gravity model can be rationalized in a variety of ways, but the various social relations added to distance to capture reductions in cost (e.g. language, common border, common FTA or currency union, etc.) are essentially ad hoc additions. Migration flows enter in the same way.

There is a growing body of work on the role of networks in trade (Rauch 2001). However, this research generally fails to distinguish between the two distinct aspects of those networks that we have already noticed: one is the role of networks in mediating the economic relationship between two dense networks; and the other is the internal structure of the networks that do the spanning. Following Ronald Burt (2005, 2009), we have referred to these as *brokerage* and *closure*.

Burt's work tends to focus on the organization of firms, but the application to broader market relations is immediate. In our context, immigrants are naturally brokers between their source country and the host country in which they settle. That is, they carry information about commodities available in source country to consumers in the host country, and vice versa. At the same time, the internal structure of a network plays a fundamental role in determining the effectiveness with which it is able to carry out the business of brokerage. Closure (i.e. relatively dense relations among members of a group) supports both reputation-building and punishment via exclusion from group benefits. Thus, especially for countries characterized by poor enforcement of contract and property rights, and/or commodities for which an arm's length market does not exist, a high degree of closure within an immigrant community permits information exchange and bonding that supports more extensive exchange.

The literature on migration and trade emphasizes the market creation (brokerage). Starting with Gould's (1994) paper that initiated the massive empirical literature using gravity models to evaluate the link between trade and migration, most of the papers in this area have seen that link in terms of spanning between dense networks of consumers.[15] The basic notion is that, in addition to their own demand for products of their home market, immigrants carry information about

[15]This body of research really is "massive". Starting from Gould's original paper until the time of writing of this paper, I count over seventy published and unpublished papers. For an extensive

those products that is useful to native consumers. Both of these will tend to increase demand for products of the immigrant home market in their new host country. Thus, both of these channels involve market creation.

In addition to seeking an accurate measure of the effect of immigration on trade, most of the empirical literature also seeks to explicitly distinguish the brokerage effect from a pure demand effect. Since, in the absence of data that distinguishes imports/exports by immigrants from imports/exports by natives, there is no straightforward structural way to make this distinction, all of these efforts involve attempts to infer which channel is relevant based on information about type of commodity, type of immigrant, or type of country involved. For example, Gould (1994) approaches this problem primarily by arguing for an asymmetry between effects on imports and exports. Specifically, he argues that pure demand effects should not have any effect on exports from the host country to the immigrant home, but should affect imports.[16] Thus, Gould's inference is that: if immigration positively affects imports, but not exports, then the demand channel is revealed to dominate; but if immigration positively affects exports, but not imports, then the brokerage effect dominates. In the event, for the case of the US 1970–1986, he finds that both are significant, but that exports are influenced more than imports by immigrant flows.[17] In addition to further studies using US data, similar studies have been done for a number of countries (e.g. Canada, UK, Australia, Switzerland, Sweden, Denmark, Spain Greece, Italy, Malaysia, and Bolivia).[18] All of these papers find a statistically significant, positive link between immigration and both imports and exports; however, there doesn't seem to be any particular pattern in the relative magnitude of the import versus the export link. Similarly, a large number of studies disaggregate the host country to subnational units (e.g. US states, Canadian Provinces, Spanish Provinces, French départements). Again, the results show significant effects of migration on trade, but no particular pattern in the effect on exports relative to imports. More recently, the development of multicountry datasets has permitted the analysis of multiple hosts and multiple homes. The great majority of these focus on (some subset of) OECD host countries and a large number of home countries (Lewer 2006; Konečný 2007a,b; Moenius et al. 2007; Dolman 2008; Lewer and Van den Berg 2008; Morgenroth and O'Brien 2008; Bettin and Turco 2010; Egger et al. 2012b; Felbermayr and Toubal 2012; Konečný 2012), while some use a matched sample of countries that trade and exchange immigrants

review of this literature, see Part I of White (2010), Chapter 2 of White and Tadesse (2011), or the meta-analysis in Genc et al. (2012).

[16]This makes sense in a partial equilibrium way. However, if the scale of either return migration or emigration of host country natives is correlated with the scale of immigration, this inference may run into trouble.

[17]As in most of the empirical literature, we use the language of immigrant flows (because that matches the theory used to interpret the results), but it should be understood that the variable in question is invariably a stock.

[18]References to the large number of papers here can be found in White and Tadesse (2011, Chapter 2) or Gaston and Nelson (2013).

(Hatzigeorgiou 2010; Parsons 2011; Tadesse and White 2011). The results here are as with the previous work on this question.

One possibility for trying to distinguish preference from information effects is to determine whether the effect of migration on trade decays above some threshold. We have already noted that both of these should increase trade. However, one might argue that demand effects should simply be linear in the immigrant population, while information effects should be subject to decay.[19] Gould (1994) did this with a non-linear functional form developed for the purpose and a handful of other papers followed suit (Head and Ries 1998; Wagner et al. 2002; Bryant et al. 2004; Morgenroth and O'Brien 2008; Egger et al. 2012b). The general result here is that the effect of migration on trade is subject to diminishing returns. Furthermore, this effect sets in at quite low levels of migration. Unless this effect is driven by the effect of migration on diffusion of preference for migrant source goods to the native population, this would seem to be strong evidence for the information link.

To further unpack this result, Gould considers consumer and producer goods separately, under the assumption that the former is more differentiated than the latter and that, as a result, the demand by immigrants will be greater. Imports and exports of both types of good are positively affected by immigration, but imports of consumer goods have the largest effect found in Gould's analysis. He takes this to suggest that both there is evidence of demand effect for consumer goods imports, but brokerage plays the dominant role in the other cases. A number of studies follow Gould in trying to find a disaggregation that provides additional leverage in distinguishing demand from brokerage effects. A variety of disaggregations are used, including: Gould's choice of consumer v. producer goods (Herander and Saavedra 2005; Mundra 2005; Blanes-Cristóbal 2008; Kandogan 2009); finished v. Intermediate (Mundra 2005); and cultural v. non-cultural goods (Tadesse and White 2008; White and Tadesse 2008; Tadesse and White 2010). In the cases of

[19]It should be noted that the econometric implications of own demand and demonstration to host natives are rather different. We would expect own demand to vary more-or-less linearly with immigration, but demonstration effects are likely to be more complex. For example, if there is a uniform propensity of natives to consume new varieties of foreign goods, and information diffuses immediately, the first immigrants provide all the relevant information, leading to an initial jump in demand, but no subsequent change other than the linear increase deriving from own demand. However, if the diffusion of information follows some specific process (or willingness to adopt does) then that process will interact with the linear immigrant process to produce some combination of the two. Most work, either implicitly or explicitly, presumes that there is a positive linear relationship between immigration and demand for imports from host countries running through the preference channel. To the extent that the information bridge runs both ways, immigrants will provide information to their home countries about host country goods, thus increasing exports and, while there is no reason that the process of learning/adoption should be the same in home and host, neither is there any reason to assume the either takes any particular form. From a welfare point of view, both of these channels should increase welfare in the context of a Krugman (1981) monopolistic competition model of the sort that underlies the Anderson-van Wincoop (2003) framework central to much of the gravity modeling used to study the empirical relationship between trade and migration. Romer (1994) makes a similar argument in his discussion of the welfare cost of trade restrictions where imports may be new goods.

all three disaggregations, the expectation is that demand effects will show up in a positive relationship between goods that are in some sense differentiated (i.e. consumer/manufactured/cultural) so that a preference for home varieties makes sense, and there is a fairly consistent pattern of immigration strongly affecting these imports.

Instead of sorting goods by some end-use category, an alternative is to sort the goods by the type of market on which they are traded. Starting from an explicitly network theoretic basis, Rauch (1999) argued that markets could be distinguished by whether: there is an organized exchange; there are reference prices quoted; and all other markets. Rauch argued that the first two types of market require that the goods traded on them be quite standardized (with organized exchanges requiring a higher degree of standardization than reference price goods) and that the residual markets, since they cannot support organized exchanges or reference prices, must be more highly differentiated. In the context of a basic gravity model, this paper found that trade costs (distance) and trade cost reduction (common language, common colonial tie) played a more significant role for differentiated goods than for the more standardized goods. In an important later paper, Rauch and Trindade (2002) argued that coethnic networks can play a significant role in reducing trade costs and, since Rauch (1999) showed that these costs are more important for differentiated goods, they should be particularly important in a gravity model as a factor increasing trade between a home and host country. Rauch and Trindade focus specifically on the Chinese diaspora (widely believed to play a major role in trade) by using Chinese population share as a variable in an otherwise standard gravity regression, finding that this variable is always significant across all types of markets, but has the greatest impact in differentiated goods—as is consistent with the hypothesis that ethnic networks of traders reduce trade costs.[20] Since the publication of Rauch and Trindade, a number of studies have used Rauch's classification in more standard setups where migrants from any country might reduce trade costs (e.g. Briant et al. 2009; Egger et al. 2012a; Felbermayr and Toubal 2012).[21] While immigrant stock tends to be a significant, positive predictor of trade (both imports and exports) in all three categories, there is no particular pattern in the magnitudes of effect (though the meta analysis in Genc et al. (2012) is consistent with a smaller effect of immigration on standardized goods).

An alternative approach reasons that emigration of host natives to the source country cannot affect imports via the preference channel, so a positive coefficient

[20]Felbermayr et al. (2010) replicate and extend the Rauch/Trindade analysis by using more current econometric techniques and by considering additional diasporas. While they estimate much smaller effects across all types of markets, they still find that the Chinese diaspora is more important for differentiated goods than for either of the standardized goods. Interestingly, they also find that, in terms of trade creation, the Moroccan, Polish, Turkish, Pakistani, Philippino, Mexican and British are all at least as important as the Chinese.

[21]More recently, a couple of studies have used a categorization based on the Broda and Weinstein (Broda and Weinstein 2006) elasticities of substitution (Tai 2009; Bettin and Turco 2010; Peri and Requena-Silvente 2010). The effects here are, if anything, weaker.

on the emigration variable should be seen as evidence for the presence of an information channel. A number of papers have checked this relationship, finding that emigration is consistently positive and significant (Canavire Bacarreza and Ehrlich 2006; Konečný 2007b; Dolman 2008; Hatzigeorgiou 2010; Parsons 2011; Felbermayr and Toubal 2012).

While the brokerage role that is emphasized by the literature is of obvious importance in understanding the link between trade and migration, closure plays a particularly central role in dealing with institutional failures and asymmetric information problems. As with brokerage, there is also a sizable literature on this second link. This work starts from the problems of contracting in certain types of goods or environments. The idea is that, in the absence of relatively complete contracts and/or effective legal environments, the risk of loss due to opportunistic behavior is sufficiently high that many mutually beneficial contracts would not be made in the absence of some alternative source of assurance. Anthropologists, sociologists and historians have long emphasized these factors in explaining the role of ethnic networks and diasporas in the organization of trade across political jurisdictions or, more generally, in the absence of effective protection of contractual/property rights (Polanyi 1957, 1968; Geertz 1963; Cohen 1969, 1971; Bonacich 1973; Geertz 1978; Curtin 1984). Recent theoretical and empirical work strongly suggests that factors such as institutional quality, business conditions, and political order constitute significant trade costs (Anderson and Marcouiller 2002; Anderson and Bandiera 2006; Anderson and Young 2006; Berkowitz et al. 2006; Turrini and Ypersele 2006; Ranjan and Lee 2007; Ranjan and Tobias 2007; Araujo et al. 2012). The earlier work by historians, anthropologists and sociologists suggest that diasporas can play a significant role in reducing these costs. Where this work identifies institutional sources of trading cost, or trading relationships that might be expected to be characterized by the presence of such costs (e.g. south-south relations, or north-south relations), it seems reasonable to expect that migration would have a particularly large trade-supporting role in the presence of such costs. Thus, a number of papers have focused specifically on developing countries as a source of both trade and migration (Co et al. 2004; White 2007; Felbermayr and Jung 2009; Bettin and Turco 2010), finding that south to north migration affects overall trade, but particularly exports of differentiated products.[22] A particularly interesting example of this sort of analysis is presented in White and Tadesse (2011, Chapters 10 & 11), where they consider four main classes of immigration corridors (north to north, north to south, south to north, and south to south), finding the strongest effect in the south to south case, no effect in the north to north case, and intermediate effects in the two north/south links. Consistent with this, analyses that include some measure of institutional quality in the source country tend to find that the trade creating effects of migrants is greater when one of the countries has poor institutional quality (Dunlevy 2006; Konečný 2007b; Briant et al. 2009).

[22]Interestingly, for the Danish case, White (2007) finds that the immigrant trade links are strongest for high income trading partners, rather than low income partners.

An interesting alternative approach to identifying this effect reasons that members of the same diaspora, on both sides of a trading dyad that does not contain the source country of the diaspora, would constitute evidence for the presence of a market-creating response to poor institutions or asymmetric information problems. Rauch and Trindade's (2002) widely cited paper, in its focus on the trade creating effect of the Chinese diaspora, is obviously an example of the application of this inference. Felbermayr et al. (2010) develop the logic of this inference in detail and implement it in a multi-source/multi-host environment.

In an interesting recent approach, a number of studies have built on Chaney's (2008) firm heterogeneity extension of Krugman's (1980) model to evaluate the effect of migration on the extensive versus the intensive margin of trade (Jiang 2007; Peri and Requena-Silvente 2010; Coughlin and Wall 2011). Peri and Requena-Silvente (for Spain) and Jiang (for Canada) find only evidence of an immigrant effect on the extensive margin, while Coughlin and Wall (using US state data) only find evidence of effects on the intensive margin. Peri and Requena-Silvente also include a product categorization based on the Broda/Weinstein elasticities and argue that their finding that immigrants affect mainly the extensive margin of exports for highly differentiated goods implies that migration reduces fixed costs, not variable costs of exporting those goods. To the extent that institutional failures and information asymmetries are interpreted as fixed costs, this interpretation of the Spanish and Canadian evidence would seem to constitute evidence for the importance of closure effects.

The work we have just discussed suggests that diasporas can play a particularly strong role in mediating trade links in the context of institutional problems and asymmetric information problems. However, much of the earlier work suggests that the internal structure of groups plays an important role in dealing with contracting problems. This is where Burt's notion of closure plays an essential role. The role of ethnic homogeneity (or, more broadly, social proximity) is also emphasized in the general literature on "middlemen minorities" and trading diasporas generally (Blalock 1967; Bonacich 1973; Iyer and Jon 1999).[23] Landa's (1981, 1994) use of social proximity in a transaction cost framework, and, while not central to the formalisms he develops, it certainly plays an essential role in the historical analysis of Greif (1989, 1991, 2006). For example, Landa's study of the role of Chinese traders in Malaysia suggests that transaction costs rise as one crosses every categorical level implying greater social distance. Greater density of members of a given degree of proximity permits more extensive division of labor within the network and, thus, a more efficient network. By focusing on very specific communities, analyses like Landa's are able to provide clear causal connections between closure and effectiveness in mediating certain types of trade relation. The downside of this approach is that one is constrained to focus on relatively small communities. It is in the nature of large-scale datasets that they do not have this

[23] An interesting specific case that has been well studied involves the role of ultra-orthodox Jews in the organization of the wholesale diamond industry (Bernstein 1992; Richman 2005, 2009).

sort of ethnographic detail. In an effort to substitute (very imperfectly) for this information, Egger et al. (2012a) attempt to use information on common origin and skill. Although this is rather blunt, they hypothesize that, by comparison to migrant communities with a wide range of skills, migrations characterized by a strong concentration of a given skill group will form more effective networks, generate "better" bridges, and thus produce a stronger link between migration and trade. This is, however, less blunt than it seems *prima facie*. After all, networks play an essential role in the location choices of emigrants. Because networks reduce the costs of migration, it is well-known that immigrants drawn from well-defined sending regions tend to go to equally well-defined locations in the receiving country.[24] Thus, when they consider the pre-existing social bonds between any group of migrants, the claim that similarity of education creates closer bonds gains considerable plausibility. This suggests a hypothesis that, other things equal, the effect of migration on trade will be stronger for migration flows made up of people with relatively homogeneous skills.[25] And this is, in fact, the main result of that paper.

6 Conclusion

Overall, then, it is clear that networks play a substantial role in virtually all aspects of the immigration experience. While the language of networks has long played a role in the analysis of migration, we are still very much just beginning to apply the formal network methods to the increasingly detailed data that are becoming available. This is clearly an area where we can predict substantial growth in new research.

[24]This is a standard of immigrant narratives, but there is a sizable body of systematic research on these links as well. One of the most compelling remains Massey et al.'s classic *Return to Aztlan* (Massey et al. 1987). In addition to a steady flow of work from the Mexican Migration Project at Princeton, directed by Massey, there is a sizable body of work across many disciplines in the social sciences (e.g. Taylor 1986; Winters et al. 2001; McKenzie and Rapoport 2007).

[25]A number of the papers on trade and migration have considered different levels of skill, but the emphasis there is on whether some level of skill is particularly strongly associated with trade creation (Hong and Santhapparaj 2006; Dolman 2008; Felbermayr and Jung 2009; Hatzigeorgiou 2010; Javorcik et al. 2011; Felbermayr and Toubal 2012). This work tends to find that skilled immigrants are strongly associated with trade creation, though intermediate levels of skill seem to have no such relationship. Closest to our work is that of Felbermayr and Toubal, which finds that share of high skilled migrants is strongly associated with exports of differentiated goods and goods traded on organized markets, but less so with goods associated with reference prices.

References

Aguilera, M. B. (2009.) Ethnic enclaves and the earnings of self-employed latinos. *Small Business Economics, 33*(4), 413–425.

Anderson, J. E. (2011.) The gravity model. *Annual Review of Economics, 3*, 133–160.

Anderson, J. E., & Bandiera, O. (2006). Traders, cops and robbers. *Journal of International Economics, 70*(1), 197–215.

Anderson, J. E., & Marcouiller, D. (2002). Insecurity and the pattern of trade: An empirical investigation. *Review of Economics and Statistics, 84*(2), 342–352.

Anderson, J. E., & VanWincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review, 93*(1), 170–192.

Anderson, J. E., & VanWincoop, E. (2004). Trade costs. *Journal of Economic Literature, 42*(3), 691–751.

Anderson, J. E., & Young, L. (2006). Trade and contract enforcement. *Contributions in Economic Analysis & Policy, 5*(1), Article 30.

Araujo, L., Mion, G., & Ornelas, E. (2012). *Instiutions and export dynamics*. CEP Discussion Paper, #1118.

Banerjee, B. (1983). Social networks in the migration process: Empirical evidence on chain migration in India. *Journal of Developing Areas, 17*(2), 185–196.

Bao, S., Bodvarsson, Ö. B., Hou, J. W., & Yaohui, Z. (2009). Migration in China from 1985 to 2000. *Chinese Economy, 42*(4), 7–28.

Bauer, T. K., Epstein, G. S., & Gang, I. N. (2007). The influence of stocks and flows on migrants location choices. *Research in Labor Economics, 26*, 199–229.

Bauer, T., & Zimmermann, K. F. (1997). Network migration of ethnic Germans. *International Migration Review, 31*(1), 143–149.

Bergstrand, J. H., & Egger, P. (2011). Gravity equations and economic frictions in the World economy. In D. M. Bernhofen, R. Falvey & U. Kreickemeier (Eds.), *Handbook of international trade* (pp. 532–570). Houndmills, Basingstoke: Palgrave Macmillan.

Berkowitz, D., Moenius, J., & Pistor, K. (2006). Trade, law, and product complexity. *Review of Economics and Statistics, 88*(2), 363–373.

Bernstein, L. (1992). Opting out of the legal system: Extralegal contractual relations in the diamond industry. *The Journal of Legal Studies, 21*(1), 115–157.

Bertoli, S. (2010). Networks, sorting and self-selection of Ecuadorian migrants. *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, (97/98), 261–288.

Bettin, G., & Turco, A. L. (2010). *A cross country view on South-North migration and trade: Dissecting the channels*. Universita' Politecnica delle Marche (I), Dipartimento di Scienze Economiche e Sociali Working Paper, #331.

Blalock, H. M. (1967). *Toward a theory of minority-group relations*. New York: Wiley.

Blanes-Cristóbal , J. V. (2008). Characteristics of immigrants and bilateral trade. *Revista De Economia Aplicada, 16*(48), 133–159.

Bonacich, E. (1973). Theory of middleman minorities. *American Sociological Review, 38*(5), 583–594.

Borjas, G. J. (1987). Self-selection and the earnings of immigrants. *American Economic Review, 77*(4), 531–553.

Borjas, G. J. (1994). Immigrant skills and ethnic spillovers. *Journal of Population Economics, 7*(2), 99–118.

Borjas, G. J. (1995). Ethnicity, neighborhoods, and human-capital externalities. *American Economic Review, 85*(3), 365–390.

Borjas, G. J. (1998). To ghetto or not to ghetto: Ethnicity and residential segregation. *Journal of Urban Economics, 44*(2), 228–253.

Borjas, G. J. (2000). Ethnic enclaves and assimilation. *Swedish Economic Policy Review, 7*(2), 89–122.

Briant, A., Combes, P.-P., & Lafourcade, M. (2009). *Product complexity, quality of institutions and the pro-trade effect of immigrants*. Paris School of Economics Working Paper, #2009-06.
Broda, C., & Weinstein, D. (2006). Globalization and the gains from variety. *Quarterly Journal of Economics, 121*(2), 541–585.
Bryant, J., Genç, M., & Law, D. (2004). *Trade and migration to New Zealand*. New Zealand Treasury, Working Paper Series, #04/18.
Burt, R. S. (1992). *Structural holes: The social structure of competition*. Cambridge, MA: Harvard University Press.
Burt, R. S. (2001). Structural holes versus network closure as social capital. In N. Lin, K. S. Cook & R. S. Burt (Eds.), *Social capital: Theory and research* (pp. 31–56). New York: Aldine de Gruyter.
Burt, R. S. (2002). The social capital of structural holes. In M. F. Guillén (Ed.), *The new economic sociology: Developments in an emerging field* (pp. 148–190). New York: Russell Sage Foundation.
Burt, R. S. (2005). *Brokerage and closure: An introduction to social capital*. Oxford: Oxford University Press.
Burt, R. S. (2009). Network duality of social capital. In V. O. Bartkus & J. H. Davis (Eds.), *Social capital: Reaching out, reaching in* (pp. 39–65). Cheltenham: Edward Elgar.
Calvo-Armengol, A., & Jackson, M. O. (2004). The effects of social networks on employment and inequality. *American Economic Review, 94*(3), 426–454.
Canavire, B., Javier, G., & Ehrlich, L. (2006). The impact of migration on foreign trade: A developing country approach. *Latin American Journal of Economic Development, 6*, 125–146.
Carrington, W. J., Detragiache, E., & Vishwanath, T. (1996). Migration with endogenous moving costs. *American Economic Review, 86*(4), 909–930.
Chaney, T. (2008). Distorted gravity: The intensive and extensive margins of international trade. *American Economic Review, 98*(4), 1707–1721.
Chau, N. H. (1997). The pattern of migration with variable migration cost. *Journal of Regional Science, 37*(1), 35–54.
Chiswick, B. R. (1978). Effect of Americanization on earnings of foreign-born men. *Journal of Political Economy, 86*(5), 897–921.
Chiswick, B. R. (1991). Speaking, reading, and earnings among low-skilled immigrants. *Journal of Labor Economics, 9*(2), 149–170.
Chiswick, B. R., & Miller, P. W. (1996). Language and earnings among immigrants in Canada: A survey. In H.O. Duleep and P.V. Wunnava (Eds.), *Immigrants and immigration policy: Individual skills, family ties, and group identities* (pp. 39–56). Greenwich, Conn.: JAI Press.
Chiswick, B. R., & Miller, P. W. (2005). Do enclaves matter in immigrant adjustment? *City and Community, 4*(1), 5–35.
Choldin, H. M. (1973). Kinship networks in the migration process. *International Migration Review, 7*(2), 163–175.
Chort, I., Gubert, F., & Senne, J.-N. (2012). Migrant networks as a basis for social control: Remittance incentives among senegalese in France and Italy. *Regional Science and Urban Economics, 42*(5), 858–874.
Co, C. Y., Euzent, P., & Martin, T. (2004). The export effect of immigration into the USA. *Applied Economics, 36*(6), 573–583.
Cohen, A. (1969). *Custom and politics in urban Africa: A study of Hausa migrants in Yoruba Towns*. London: Routledge & Kegan Paul.
Cohen, A. (1971). Cultural strategies in the organization of trading diasporas. In C. Meillassoux (Ed.), *The development of indigenous trade and markets in West Africa* (pp. 266–278). Oxford: Oxford University Press.
Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology, 94*(Supplement), S95–S120.
Coughlin, C. C., & Wall, H. J. (2011). Ethnic networks and trade: Intensive versus extensive margins." *Economics Letters, 113*(1), 73–75.

Curtin, P. D. (1984). *Cross-cultural trade in World history*. Cambridge, Cambridgeshire/New York: Cambridge University Press.

Cutler, D. M., & Glaeser, E. L. (1997). Are ghettos good or bad? *Quarterly Journal of Economics, 112*(3), 827–872.

Cutler, D. M., Glaeser, E. L., & Vigdor, J. L. (2008). When are Ghettos bad? Lessons from immigrant segregation in the United States. *Journal of Urban Economics, 63*(3), 759–774.

Damm, A. P. (2009). Ethnic enclaves and immigrant labor market outcomes: Quasi-experimental evidence. *Journal of Labor Economics, 27*(2), 281–314.

Danzer, A. M., & Yaman, F. (2013). Do ethnic enclaves impede immigrants' integration? Evidence from a quasi-experimental social-interaction approach. *Review of International Economics, 21*(2), 311–325.

Davis, B., Stecklov, G. U. Y., & Winters, P. (2002). Domestic and international migration from rural Mexico: Disaggregating the effects of network structure and composition. *Population Studies, 56*(3), 291–309.

Deléchat, C. (2001). International migration dynamics: The role of experience and social networks. *Labour, 15*(3), 457–486.

Dolfin, S., & Genicot, G. (2010). What do networks do? The role of networks on migration and 'Coyote' use. *Review of Development Economics, 14*(2), 343–359.

Dolman, B. (2008). *Migration, trade and investment*. Productivity Commission Staff Working Paper

Dunlevy, J. A. (2006). The influence of corruption and language on the protrade effect of immigrants: Evidence from the American states. *Review of Economics and Statistics, 88*(1), 182–186.

Durand, J., & Massey, D. S. (2004). *Crossing the border: Research from the Mexican migration project*. New York: Russell Sage Foundation.

Dustmann, C. (1994). Speaking fluency, writing fluency and earnings of migrants. *Journal of Population Economics, 7*(2), 133–156.

Dustmann, C., & van Soest, A. (2001). Language fluency and earnings: Estimation with misclassified language indicators. *Review of Economics and Statistics, 83*(4), 663–674.

Edin, P.-A., Fredriksson, P., & Åslund, O. (2003). Ethnic enclaves and the economic success of immigrants: Evidence from a natural experiment. *Quarterly Journal of Economics, 118*(1), 329–357.

Egger, P. H., Von Ehrlich, M., & Nelson, D. R. (2012a). *The trade effects of skilled versus unskilled migration*. CESifo Working Paper Series, #9053.

Egger, P. H., von Ehrlich, M., & Nelson, D. R. (2012b). Migration and trade. *The world economy, 35*(2), 216–241.

Felbermayr, G. J., & Jung, B. (2009). The pro-trade effect of the brain drain: Sorting out confounding factors. *Economics Letters, 104*(2), 72–75.

Felbermayr, G. J., & Toubal, F. (2012). Revisiting the trade-migration nexus: Evidence from new OECD data. *World Development, 40*(5), 928–937.

Felbermayr, G., Jung, B., & Toubal, F. (2010). Ethnic networks, information, and international trade: Revisiting the evidence. *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, (97/98), 41–70.

Flores-Yeffal, N. Y., & Aysa-Lastra, M. (2011). Place of origin, types of ties, and support networks in Mexico-U.S. migration. *Rural Sociology, 76*(4), 481–510.

Flores-Yeffal, N. Y. (2013). *Migration-trust networks: Social cohesion in Mexican US-bound emigration*. College Station: Texas A&M University Press.

Fussell, E., Massey, D. S. (2004). The limits to cumulative causation: International migration from Mexican urban areas. *Demography, 41*(1), 151–171.

Fussell, E. (2004). Sources of Mexico's migration stream: Rural, urban, and border migrants to the United States. *Social Forces, 82*(3), 937–967.

Garip, F. (2008). Social capital and migration: How do similar resources lead to divergent outcomes? *Demography, 45*(3), 591–617.

Gaston, N., & Nelson, D. R. (2013). Bridging trade theory and labour econometrics: The effects of international migration. *Journal of Economic Surveys, 27*(1), 98–139.
Geertz, C. (1963). *Peddlers and princes; social change and economic modernization in two Indonesian towns*. Chicago: University of Chicago Press.
Geertz, C. (1978). Bazaar economy: Information and search in peasant marketing. *American Economic Review, 68*(2), 28–32.
Genc, M., Gheasi, M., Nijkamp, P., & Poot, J. (2012). The impact of immigration on international trade: A meta-analysis. In P. Nijkamp, J. Poot & M. Sahin (Eds.), *Migration impact assessment* (pp. 301–337). Cheltenham: Edward Elgar.
Gould, D. M. (1994). Immigrant links to the home country: Empirical implications for United States bilateral trade flows. *Review of Economics and Statistics, 76*(2), 302–316.
Greif, A. (1989). Reputation and coalitions in medieval trade: Evidence on the maghribi traders. *Journal of Economic History, 49*(4), 857–882.
Greif, A. (1991). The organization of long-distance trade: Reputation and coalitions in the Geniza documents and Genoa during the 11th-century and 12th-century. *Journal of Economic History, 51*(2), 459–462.
Greif, A. (2006). *Institutions and the path to the modern economy: Lessons from medieval trade*. Cambridge: Cambridge University Press.
Habermas, J. (1975). *Legitimation crisis*. Boston: Beacon Press.
Hamilton, B., & Whalley, J. (1984). Efficiency and distributional implications of global restrictions on labor mobility—calculations and policy implications. *Journal of Development Economics, 14*(1–2), 61–75.
Harris, J. M., Hirst, J. L., & Mossinghoff, M. J. (2008). *Combinatorics and graph theory*. New York: Springer.
Hatzigeorgiou, A. (2010). Migration as trade facilitation: Assessing the links between international trade and migration. *The B.E. Journal of Economic Analysis & Policy, 10*(1), Article 24.
Haug, S. (2008). Migration networks and migration decision-making. *Journal of Ethnic and Migration Studies, 34*(4), 585–605.
Head, K., & John R. (1998). Immigration and trade creation: Econometric evidence from Canada. *Canadian Journal of Economics, 31*(1), 47–62.
Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153–161.
Heckman, J. J., & Honore, B. E. (1990). The empirical content of the Roy model. *Econometrica, 58*(5), 1121–1149.
Herander, M. G., & Saavedra, L. A. (2005). Exports and the structure of immigrant-based networks: The role of geographic proximity. *Review of Economics and Statistics, 87*(2), 323–335.
Hicks, J. (1932). *The theory of wages*. London: Macmillan.
Hong, T. C., & Santhapparaj, A. S. (2006). Skilled labor immigration and external trade in Malaysia: A pooled data analysis. *Perspectives on Global Development and Technology, 5*(4), 351–366.
Ioannides, Y. M., & Loury, L. D. (2004). Job information networks, neighborhood effects, and inequality. *Journal of Economic Literature, 42*(4), 1056–1093.
Iyer, G. R., & Jon, M. S. (1999). Ethnic entrepreneurial and marketing systems: Implications for the global economy. *Journal of International Marketing, 7*(4), 83–110.
Javorcik, B. S., Ozden, C., Spatareanu, M., & Neagu, C. (2011). Migrant networks and foreign direct investment. *Journal of Development Economics, 94*(2), 231–241.
Jayet, H., Ukrayinchuk, N., De Arcangelis, G. (2010). The location of immigrants in Italy: Disentangling networks and local effects. *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, (97/98), 329–350.
Jiang, S. (2007). *Immigration, information and trade margins*. Department of Economics, University of Calgary Working Paper, #2007-16.
Jones, M. A. (1992). *American immigration*. Chicago: University of Chicago Press.

Kanas, A., Chiswick, B. R., van der Lippe, T., & van Tubergen, F. (2012). Social contacts and the economic performance of immigrants: A panel study of immigrants in Germany. *International Migration Review, 46*(3), 680–709.

Kandogan, Y. (2009). Immigrants, cross-cultural communication and export performance: The Swiss case. *European Journal of International Management, 3*(3), 393–410.

Kirman, A. P. (2011). *Complex economics: Individual and collective rationality*. London/New York: Routledge.

Koneény, T. (2007a). *Can immigrants hurt trade*? CERGEI-EJ working paper, #329.

Koneény, T. (2007b). *Immigrant links, trade creation and trade diversion*. CERGEI-EJ working paper, #329.

Koneény, T. (2012). Expatriates and trade. *Journal of International Migration and Integration, 13*(1), 83–98.

Kritz, M. M., Lim, L. L., & Zlotnik, H. (1992). *International migration systems: A global approach*. Oxford: Clarendon Press.

Krugman, P. (1980). Scale economies, product differentiation, and the pattern of trade. *The American Economic Review, 70*(5), 950–959.

Krugman, P. R. (1981). Intraindustry specialization and the gains from trade. *Journal of Political Economy, 89*(5), 959–973.

Landa, J. T. (1981). A theory of the ethnically homogeneous middleman group: An institutional alternative to contract law. *Journal of Legal Studies, 10*(2), 348–362.

Landa, J. T. (1994). *Trust, ethnicity, and identity: Beyond the new institutional economics of ethnic trading networks, contract law, and gift-exchange*. Ann Arbor: University of Michigan Press.

Lazear, E. P. (1999). Culture and language. *Journal of Political Economy, 107*(6), S95–S126.

Lazear, E. P. (2007). Mexican assimilation in the United States. In G. J. Borjas (Ed.), *Mexican immigration to the United States* (pp. 107–122). Chicago: University of Chicago Press/NBER.

Levy, M. B., Wadycki, W. J. (1973). The influence of family and friends on geographic labor mobility: An international comparison. *The Review of Economics and Statistics, 55*(2), 198–203.

Lewer, J. J. (2006). The impact of migration on bi-lateral trade: OECD results from 1991–2000. *Southwestern Economic Review, 33*(1), 187–230.

Lewer, J. J., & Van den Berg, H. (2008). A gravity model of immigration. *Economics Letters, 99*(1), 164–167.

Light, I., Sabagh, G., Bozorgmehr, M., & Der-Martirosian, C. (1994). Beyond the ethnic enclave economy. *Social Problems, 41*(1), 65–80.

Lin, N. (2001). *Social capital: A theory of social structure and action*. New York: Cambridge University Press.

Mahuteau, S., & Junankar, P. N. (2008). Do migrants get good jobs in Australia? The role of ethnic networks in job search. *Economic Record, 84*, S115–130.

Marques, H. (2010). Migration creation and diversion in the European Union: Is Central and Eastern Europe a 'Natural' member of the single market for labour? *Jcms-Journal of Common Market Studies, 48*(2), 265–291.

Massey, D. S. (1990). Social structure, household strategies, and the cumulative causation of migration. *Population Index, 56*(1), 3–26.

Massey, D. S., Alarcón, R., Durand, J., & Gonzalez, H. (1987). *Return to Aztlan: The social process of international migration from Western Mexico*. Berkeley: University of California Press.

Massey, D. S., & Aysa-Lastra, M. (2011). Social capital and international migration from Latin America. *International Journal of Population Research*, vol. 2011, Article ID 834145, 18 pages, 2011. doi:10.1155/2011/834145

Massey, D. S., & Espinosa, K. E. (1997). What's driving Mexico-US migration? A theoretical, empirical, and policy analysis. *American Journal of Sociology, 102*(4), 939–999.

McKenzie, D., & Rapoport, H. (2007). Network effects and the dynamics of migration and inequality: Theory and evidence from Mexico. *Journal of Development Economics, 84*(1), 1–24.

McKenzie, D., & Rapoport, H. (2010). Self-selection patterns in Mexico-U.S. migration: The role of migration networks. *Review of Economics and Statistics, 92*(4), 811–821.
Mincer, J. (1978). Family migration decisions. *Journal of Political Economy, 86*(5), 749–773.
Moenius, J., Rauch, J. E., & Trindade, V. (2007). Gravity and matching. San Diego: University of California.
Morgenroth, E., & O'Brien, M. (2008). *Some further results on the impact of migrants on trade*. DYNREG Working Paper, #26/2008.
Mundra, K. (2005). Immigration and international trade: A semiparametric empirical investigation. *Journal of International Trade & Economic Development, 14*(1), 65–91.
Munshi, K. (2003). Networks in the modern economy: Mexican migrants in the US labor market. *Quarterly Journal of Economics, 118*(2), 549–599.
Munshi, K. (2011). Strength in numbers: Networks as a solution to occupational traps. *The Review of Economic Studies, 78*(3), 1069–1101.
Nannestad, P., Svendsen, G. L. H., & Svendsen, G. T. (2008). Bridge over troubled water? Migration and social capital. *Journal of Ethnic & Migration Studies, 34*(4), 607–631.
Orrenius, P. M., & Zavodny, M. (2005). Self-selection among undocumented immigrants from Mexico. *Journal of Development Economics, 78*(1), 215–240.
Palloni, A., Massey, D. S., Ceballos, M., Espinosa, K., & Spittel, M. (2001). Social capital and international migration: A test using information on family networks. *American Journal of Sociology, 106*(5), 1262–1298.
Parsons, C. R. (2011). Do migrants really foster trade? The trade-migration nexus, a panel approach 1960–2000. *World Bank Policy Research Working paper*, #6034. The World Bank, Washington, DC.
Pedersen, P. J., Pytlikova, M., & Smith, N. (2008). Selection and network effects: Migration flows into OECD countries 1990–2000. *European Economic Review, 52*(7), 1160–1186.
Peri, G., & Requena-Silvente, F. (2010). The trade creation effect of immigrants: Evidence from the remarkable case of Spain. *Canadian Journal of Economics, 43*(4), 1433–1459.
Phillips, J. A., & Massey, D. S. (2000). Engines of immigration: Stocks of human and social capital in Mexico. *Social Science Quarterly, 81*(1), 33–48.
Polanyi, K. (1944/2001). *The great transformation: The political and economic origins of our time*. Boston, MA: Beacon Press.
Polanyi, K. (1957). *Trade and market in the early empires; economies in history and theory*. Glencoe, IL: Free Press.
Polanyi, K. (1968). *Primitive, archaic, and modern economies; essays of Karl Polanyi*. Garden City, NY: Anchor Books.
Portes, A., & Jensen, L. (1989). The enclave and the entrants: Patterns of ethnic enterprise in Miami before and after mariel. *American Sociological Review, 54*(6), 929–949.
Portes, A., & Rumbaut, R. G. (2006). *Immigrant America: A portrait*. Berkeley: University of California Press.
Portes, A., & Sensenbrenner, J. (1993). Embeddedness and immigration: Notes on the social determinants of economic action. *American Journal of Sociology, 98*(6), 1320–1350.
Portes, A., & Stepick, A. (1993). *City on the edge: The transformation of Miami*. Berkeley: University of California Press.
Portes, A., & Zhou, M. (1996). Self-employment and the earnings of immigrants. *American Sociological Review, 61*(2), 219–230.
Portes, A., & Zhou, M. (1999). Entrepreneurship and economic progress in the 1990s: A comparative analysis of immigrants and African Americans. In F. D. Bean & S. Bell-Rose (Eds.), *Immigration and opportunity: Race, ethnicity, and employment in the United States* (pp. 143–171). New York: Russell Sage Foundation.
Ranjan, P., & Lee, J. Y. (2007). Contract enforcement and international trade. *Economics & Politics, 19*(2), 191–218.
Ranjan, P., & Tobias, J. L. (2007). Bayesian inference for the gravity model. *Journal of Applied Econometrics, 22*(4), 817–838.

Rauch, J. E. (1999). Networks versus markets in international trade. *Journal of International Economics, 48*(1), 7–35.

Rauch, J. E. (2001). Business and social networks in international trade. *Journal of Economic Literature, 39*(4), 1177–1203.

Rauch, J. E., & Trindade, V. (2002). Ethnic Chinese networks in international trade. *Review of Economics and Statistics, 84*(1), 116–130.

Richman, B. D. (2005). How community institutions create economic advantage: Jewish diamond merchants in New York. *American Law & Economics Association Annual Meetings* (Vol. 51). Bepress, Berkeley, California.

Richman, B. D. (2009). Ethnic networks, extralegal certainty, and globalisation: Peering into the diamond industry. In V. Gessner (Ed.), *Legal certainty beyond the state*. Oxford: Hart Publishing.

Romer, P. (1994). New goods, old theory, and the welfare costs of trade restrictions. *Journal of Development Economics, 43*(1), 5–38.

Sanders, J. M., & Nee, V. (1987). Limits of ethnic solidarity in the enclave economy. *American Sociological Review, 52*(6), 745–773.

Sanders, J. M., & Nee, V. (1992). Problems in resolving the enclave economy debate. *American Sociological Review, 57*(3), 415–418.

Schumpeter, J. A. (1942/1975). *Capitalism, socialism, and democracy*. New York: Harper & Row.

Sjaastad, L. A. (1962). The costs and returns of human migration. *Journal of Political Economy, 70*(5), 80–93.

Solow, R. M. (2000). Notes on social capital and economic performance. In P. Dasgupta & I. Serageldin (Eds.), *Social capital: A multifaceted perspective* (pp. 6–10). Washington, DC: The World Bank.

Spilimbergo, A., & Ubeda, L. (2004). A model of multiple equilibria in geographic labor mobility. *Journal of Development Economics, 73*(1), 107–123.

Stark, O., & Jakubek, M. (2013). Migration networks as a response to financial constraints: Onset, and endogenous dynamics. *Journal of Development Economics, 101*(0), 1–7.

Stark, O., & Levhari, D. (1982). On migration and risk in ldcs. *Economic Development and Cultural Change, 31*(1), 191–196.

Tadesse, B., & White, R. (2008). Do immigrants counter the effect of cultural distance on trade? Evidence from US state-level exports. *Journal of Socio-Economics, 37*(6), 2304–2318.

Tadesse, B., & White, R. (2010). Cultural distance as a determinant of bilateral trade flows: Do immigrants counter the effect of cultural differences? *Applied Economics Letters, 17*(2), 147–152.

Tadesse, B., & White, R. (2011). Emigrant effects on trade: Re-examining the immigrant-trade link from the home country perspective. *Eastern Economic Journal, 37*(2), 281–302.

Tai, S. H. T. (2009). Market structure and the link between migration and trade. *Review of World Economics, 145*(2), 225–249.

Taylor, J. E. (1986). Differential migration, networks, information, and risk. In O. Stark (Ed.), *Migration, human capital and development, research in human capital and development* (pp. 141–173). New York: JAI Press.

Teteryatnikova, M. (2013). A Model of Social Networks and Migration Decisions. Vienna, MS: University of Vienna.

Tunali, I. (2000). Rationality of migration. *International Economic Review, 41*(4), 893–920.

Turrini, A. A. & van Ypersele, T. (2006). *Legal costs as barriers to trade*. CEPR Discussion Paper, #5751.

Wagner, D., Head, K., & Ries, J. (2002). Immigration and the trade of provinces. *Scottish Journal of Political Economy, 49*(5), 507–525.

Walmsley, T. L., Winters, A., & Ahmed. A. (2011). The impact of the movement of labour: Results from a model of bilateral migration flows. *Global Economy Journal, 11*(4) 22, doi:10.2202/1524-5861.1738.

Warman, C. (2007). Ethnic enclaves and immigrant earnings growth. *The Canadian Journal of Economics, 40*(2), 401–422.

Wegge, S. A. (1998). Chain migration and information networks: Evidence from Nineteenth-century Hesse-Cassel. *Journal of Economic History, 58*(4), 957–986.
White, R. (2007). An examination of the Danish immigrant-trade link. *International Migration, 45*(5), 61–86.
White, R. (2010). *Migration and international trade : The US experience since 1945*. Cheltenham: Edward Elgar.
White, R., & Tadesse, B. (2008). Cultural distance and the US Immigrant-trade link. *World Economy, 31*(8), 1078–1096.
White, R., & Tadesse, B. (2011). *International migration and economic integration: Understanding the immigrant-trade link*. Cheltenham/Northampton, MA: Edward Elgar.
Wilson, K. L., & Portes, A. (1980). Immigrant enclaves: An analysis of the labor market experiences of Cubans in Miami. *American Journal of Sociology, 86*(2), 295–319.
Winters, P., de Janvry, A., & Sadoulet, E. (2001). Family and community networks in Mexico-US migration. *Journal of Human Resources, 36*(1), 159–184.
Wong, S. L., & Salaff, J. W. (1998). Network capital: Emigration from Hong Kong. *British Journal of Sociology, 49*(3), 358–374.
Xie, Y., & Gough, M. (2011). Ethnic enclaves and the earnings of immigrants. *Demography, 48*(4), 1293–1315.
Yang, D. (2008). International migration, remittances, and household investment: Evidence from Philippine migrants' exchange rate shocks. *Economic Journal, 118*(528), 591–630.
Yang, D. (2011). Migrant remittances. *Journal of Economic Perspectives, 25*(3), 129–152.
Yang, D. & HwaJung C. (2007). Are remittances insurance? Evidence from rainfall shocks in the Philippines. *The World Bank Economic Review, 21*(2), 219–248.
Zahniser, S. (1999). *Mexican migration to the United States : The role of migration networks and human capital accumulation*. New York: Garland Pub.
Zhao, Y. (2003). The role of migrant networks in labor migration: The case of China. *Contemporary Economic Policy, 21*(4), 500–511.

The Response of German Establishments to the 2008–2009 Economic Crisis

Lutz Bellmann, Hans-Dieter Gerner, and Richard Upward

Abstract The global economic and financial crisis which began in 2008 had very different effects on the labour markets of EU economies. In particular, unemployment rose far more in some countries than in others, even after conditioning on the fall in GDP. Thus, in the words of the OECD Employment Outlook (2012), some labour markets might be described as more "resilient" than others in the face of shocks. In this chapter we propose a simple descriptive methodology which relates output shocks and job flows to hires and separations. This methodology sheds light on many of the proposed explanations for the resilience of German establishments to the crisis, in particular the role of various institutional arrangements intended to promote workplace flexibility, such as short-time-work and working time accounts. The increasing availability of detailed linked employer-employee data will enable this methodology to be applied consistently across countries. The chapter therefore serves to open up a research agenda which compares the behaviour of firms' hiring and firing policies across countries.

1 Introduction

It has been long been recognised that country specific labour market institutions are an important determinant of labour market outcomes (see Boeri and van Ours 2013, for a recent example of this view). A striking feature of the 2008–2009 economic

L. Bellmann
Friedrich-Alexander-Universität Erlangen-Nürnberg and Institut für Arbeitsmarkt-und Berufsforschung, Regensburger Straße 104, 90478 Nürnberg, Germany
e-mail: lutz.bellmann@iab.de

H.-D. Gerner
Institut für Arbeitsmarkt-und Berufsforschung, Regensburger Straße 104, 90478 Nürnberg, Germany
e-mail: hans-dieter.gerner@iab.de

R. Upward (✉)
School of Economics, University of Nottingham, Clive Granger Building, Nottingham NG7 2RD, UK
e-mail: richard.upward@nottingham.ac.uk

P. Commendatore et al. (eds.), *Complexity and Geographical Economics*,
Dynamic Modeling and Econometrics in Economics and Finance 19,
DOI 10.1007/978-3-319-12805-4_8

crisis is the extent to which different labour markets diverged in their response, and indeed the crisis may well have slowed or even stopped the process of convergence between EU countries. An open question is then whether certain labour market institutions and policies allowed some countries to "weather the storm".

In this chapter we focus on the German labour market which has, so far, shown remarkable resilience in the face of the economic crisis. Several recent articles (discussed more fully in Sect. 2 below) note that the response of the German labour market has been "astonishingly mild" Möller (2010), even though Germany experienced one of the strongest declines in GDP amongst the industrialised economies.[1] The German labour market is generally regarded as having one of the highest levels of protection against worker dismissal: the OECD's employment protection index (Venn 2009) ranks Germany 34th out of 40 countries for the protection of permanent workers against dismissal, and 36th for the requirements for collective dismissals. In addition to these pre-existing features, Germany also made use of policies such as short-time work and flexible working time in response to the crisis.

We will examine the response of a large panel of German establishments to the crisis (measured at the establishment level) in terms of their job flows (changes in employment) and the consequent worker flows (hires, separations and layoffs).[2] The proposed methodology relies on the use of linked data on employers and employees. The increasing availability of such data for many EU countries (see for example the review in Abowd and Kramarz 1999) implies that the methodology may be consistently applied across countries to examine whether the job and worker flow response of employers differs across countries in systematic ways which are related to their labour market institutions. The chapter therefore serves to open up a research agenda which compares the behaviour of firms' hiring and firing policies across countries.

Figure 1 tells the basic story. The unemployment rate in the recent recession (left-hand panel) barely increased, and is still some 4 % points *lower* than in 2005, at the end of the last downturn. The right-hand panel shows that, as GDP shrank quite dramatically in the 2008–2009 crisis, employment held up while average hours fell. Thus, overall, German firms reacted to the crisis by adjusting on the intensive margin (hours per worker) rather than on the extensive margin (number of workers).

A key point to note is that the crisis was particularly serious in the manufacturing sector in Germany, in contrast to many other OECD countries. This is illustrated in Fig. 2, which shows that the fall in output and the fall in hours was significantly larger in the manufacturing sector than in the economy as a whole. Remarkably, the fall in employment in manufacturing was only slightly greater than in the economy

[1] OECD (2012, Chapter 2) discusses in detail the extent of labour market resilience of OECD labour markets.

[2] Analysis of the intensive margin (hours of work) is more problematic because the survey we use only measures "standard" hours of work, but we will also consider within-establishment changes in labour productivity.

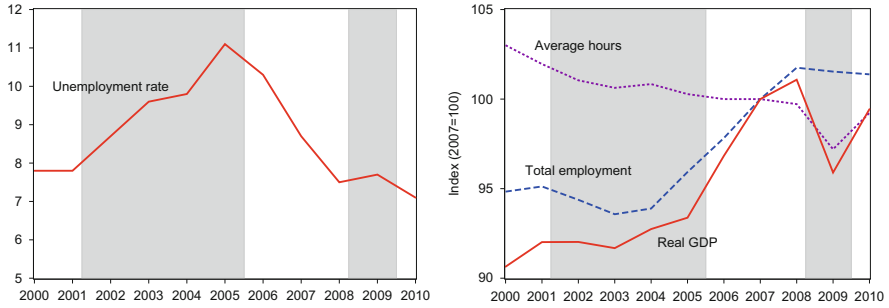


Fig. 1 German unemployment rate, real GDP, total employment and average hours 2000–2010. GDP, employment and hours indexed 2007=100. *Source*: OECD StatExtracts. Recession dates indicated by *shaded areas* are those used by Burda and Hunt (2011), originally from *Sachverständigenrat* (2010) and cover Q1 2001–Q2 2005 and Q1 2008–Q2 2009

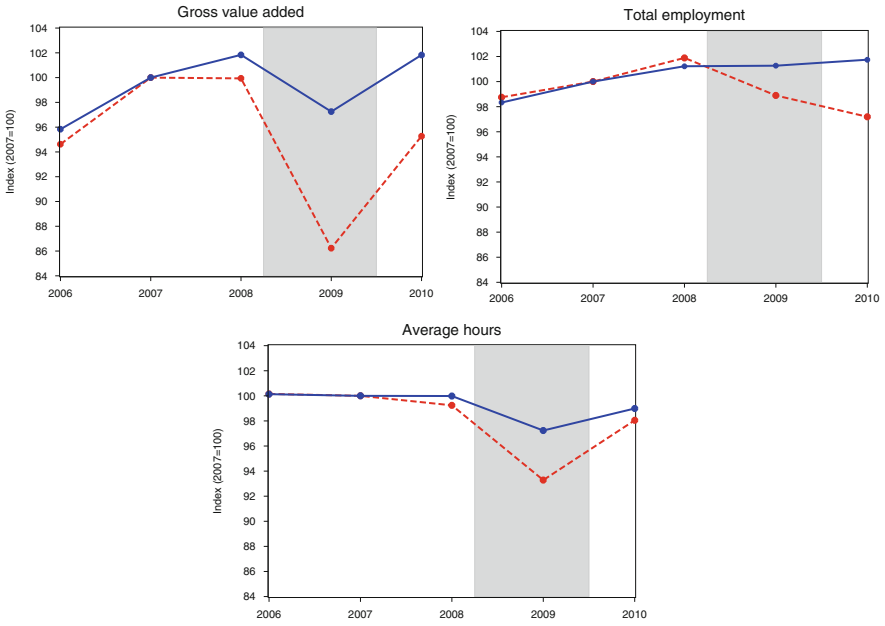


Fig. 2 Differences between manufacturing and whole economy 2006–2010. GVA, employment and hours indexed 2007-100. *Source*: German Federal Employment Agency

as a whole, although of concern is the fact that employment in manufacturing has not rebounded in 2010.

The chapter is structured as follows: in Sect. 2 we summarise a number of recent studies which have proposed a variety of explanations for the resilience of the German labour market. In Sect. 3 we describe the data we use, which comes from the IAB establishment panel survey. In Sect. 4 we show that the IAB

establishment panel does capture the key elements of the crisis: namely, a much larger fall in output than in employment. We also present descriptive evidence on how establishments reacted to the crisis in terms of employment, output, separations and layoffs. In Sect. 5 we show how employment changes in German establishments can be decomposed into hiring and layoffs, and we describe the basic short-term relationship between demand shocks and employment change. In Sect. 6 we compare the use of labour market policies (such as short-time work) between crisis and non-crisis plants, and we evaluate the effectiveness of these policies. In Sect. 7 we present some preliminary evidence on the recovery period using the latest available wave of the panel survey (2010). Section 8 concludes.

2 Explanations for the German Labour Market "Miracle"

A number of explanations have been suggested for the resilience of the German labour market. All these explanations have in common the idea that German firms adjusted their labour input at the *intensive margin* (hours of work) rather than at the *extensive margin* (number of workers). This is often referred to as "labour hoarding", although one should distinguish between falls in output per hour worked and falls in output per worker. A fall in output can be accommodated through three different channels. First, firms can adjust at the extensive margin by reducing the number of workers. Second, firms can adjust at the intensive margin by reducing the number of hours per worker. Third, output per worker-hour may fall.[3] Some authors refer to labour hoarding as being both adjustment in hours and adjustment in output per hour. But if reductions in hours lead to a similar reduction in the wage bill, only the third channel actually represents hoarding.

To be convincing, any explanation has to show why German firms used labour hoarding in this crisis, but not in previous downturns, and it has to show why German firms reacted differently to firms in other countries where employment fell much more sharply in the downturn. In this section we discuss the evidence for and against each of the proposed explanations.

2.1 The Use of Short-Time Work (STW)

Perhaps the most widely-discussed explanation for the use of labour hoarding (rather than large falls in employment) is the use of short-time work (*Kurzarbeit*.) The short-time working scheme has existed in Germany since early in the twentieth century (see Brenke et al. 2011 for a brief history). Under this measure, employers may reduce working time of their employees if they face a documented shortfall of

[3] Burda and Hunt (2011, Equation (1)) formalises these three channels.

demand.[4] Employees' loss in income is compensated (by the firm) for between 60 and 67 % of the difference between their net income before and after the working time reduction, and the firm is subsequently compensated by the German Federal Employment Agency. Originally, employers were also required to pay the full social security contribution based on employees' income before the cut in working time. However, during the 2008–2009 crisis the government paid up to half of the social security contributions. In addition, if employers combined short-time work with further training, the Federal Employment agency also paid the full social security contributions for the difference in the wages before and after the working time reduction. The maximum period of eligibility was extended from 6 to 18 months in Autumn 2008 and to 24 months in July 2009. The maximum duration was subsequently reduced again to 12 months (until at least the end of 2011).

In Fig. 3 we plot the numbers of establishments and workers using STW from 1991 onwards. Note that in the early 1990s the use of STW was dominated by plants and workers in East Germany using the so-called transfer Kurzarbeit. The mild recession of the early 2000s saw only a small increase in the use of STW. In contrast, by the end of 2009 almost 80,000 establishments were using STW, affecting about 1.5 million employees. Take-up of STW in Germany rose from less than 0.1 % of employment in 2007 to over 4 % in the second quarter of 2009 (Hijzen and Venn 2011, Figure 8).

Almost by definition, STW schemes encourage firms to use the intensive rather than the extensive margin as a response to a fall in demand. However, in an accounting sense, Möller (2010, Table 5) shows that the use of STW accounts for only a proportion of the reduction in total labour input; reductions in overtime and working time accounts were almost as important, and reductions in productivity

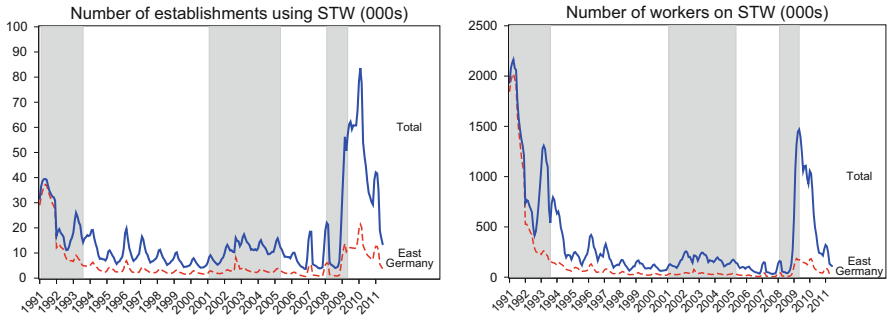


Fig. 3 Use of Kurzarbeit by establishments and workers 1991–2011. *Source*: Bundesagentur für Arbeit

[4] As explained in Crimmann et al. (2010), there are three types of STW: "transfer kurzarbeit" which was used extensively during reunification; "seasonal kurzarbeit" and "short-time work for economic reasons". We consider only the third type here.

per hour were more important.[5] In addition, both Boysen-Hogrefe and Groll (2010, Figure 1) and Burda and Hunt (2011, Figure 9) show that STW was used heavily in the recessions of the mid 1970s and early 1980s, and in those recessions employment fell far more than in the most recent crisis. Thus, on its own, STW cannot account for the German success story.

The key difficulty in evaluating the effectiveness of STW in encouraging adjustment at the intensive rather than the extensive margin is one of selection. Firms that use the STW program do so (presumably) because they face larger negative demand shocks than firms which do not use STW. Thus, comparing the employment growth of STW with non-STW firms is likely to seriously underestimate the impact of the program. Indeed, we will show in Sect. 6.2 that STW firms have worse employment performance than non-STW firms in the crisis period. In principle, one can include control variables which capture the extent of the demand shock (for example the change in sales), but in practice there are likely to be unobserved firm-specific shocks which mean that such a comparison is still flawed.

One solution is to use cross-country evidence and rely on the assumption that the availability and take-up of STW across countries is not correlated with the shock to labour demand. Early examples of this cross-country approach are Abraham and Houseman (1994) and Van Audenrode (1994). Abraham and Houseman (1994) argue that there is greater labour market adjustment on the intensive margin in European countries (France, Belgium and Germany) because of a combination of job security regulations with worksharing. Van Audenrode (1994) analyses hours and employment adjustment in ten OECD countries, and finds that working-time adjustments tend to compensate for labour force adjustment in countries with more generous short-time working arrangements.

Recent cross-country studies on the effectiveness of STW include Hijzen and Venn (2011), Cahuc and Carcillo (2011) and Arpaia et al. (2010). Hijzen and Venn's (2011) approach is to compare the employment adjustment of countries before and after the crisis according to the intensity of their use of STW. This difference-in-difference approach is intended to control for pre-existing differences between countries (such as labour market regulations). Their estimates suggest that STW schemes had an "economically important impact on preserving jobs during the economic downturn", particularly in Germany and Japan. Arpaia et al. (2010) is a similar cross-country study of STW. A difference-in-difference comparison of countries which had STW schemes in 2007 shows that these countries had smaller falls in employment during the crisis.

Cahuc and Carcillo (2011) argue that the selection problem is also relevant to cross-country studies, because the availability of STW may well be correlated with the severity of the crisis. Indeed, as we have noted, in Germany the rules governing access to the program were loosened in 2008 and 2009. Cahuc and Carcillo (2011) show that unemployment increased more in countries with higher

[5] Although note that accounting exercises such as this one do not necessarily tell us about the effectiveness of the policy because they ignore deadweight and displacement effects.

STW take-up rates, which, they argue, reflects the endogeneity of the take-up rate. They therefore instrument the use of STW with features of the program *before* the crisis. The argument is that differences in the STW program before the crisis should be uncorrelated with the strength of the demand shock, but should still be strongly correlated with the use of the program during the crisis. The results of an IV regression suggest that the use of STW had a significant negative effect on unemployment rates during the crisis. However, note that, at the country level, there may be another potential selection problem because the use of STW policies may be correlated with other institutional features which are associated with labour hoarding.

Boeri and Bruecker (2011) analyse the effectiveness of STW using both a macro (cross-country) and a micro approach. Boeri and Bruecker instrument the use of STW and working-time accounts in 2009 with establishments' use of such policies in earlier periods before the 2008–2009 crisis. This instrument is plausibly exogenous if the 2009 shock was independent of earlier shocks. In other words, an establishment which uses STW in, say, 2003, is more likely to use STW in 2009, but the earlier use of STW is not correlated with the size of the demand shock in 2009. They find that the use of STW has an economically sizable positive impact on employment growth rates between 2008 and 2009, and that this effect is much larger when the use of STW is instrumented, as one would expect given the direction of the selection effect.

To summarise, there is a considerable body of cross-country evidence from both before and during the crisis to suggest that the use of STW can significantly increase adjustment at the intensive margin, and that this can reduce the unemployment effects of a recession. There is much less evidence at the firm-level for Germany, but what there is also supports the idea that STW can prevent job loss. However, it is also clear that STW on its own cannot account for the German labour market miracle. A simple decomposition of hours changes shows that the number of hours lost through STW is far too small to account for the total loss of hours in Germany in 2009. Furthermore, the uptake of STW is no greater than in earlier downturns, when unemployment increases were far greater.

2.2 Working-Time Accounts (WTA)

Working time accounts (*Arbeitszeitkonten*) are firm-level agreements which allow actual working hours to vary from agreed working hours within defined limits. WTAs also specify the period over which compensation of working time must occur; this is most commonly 1 year (Seifert 2005), but may be longer or shorter. Total pay does not vary with actual hours worked, so in effect hourly wage rates vary inversely with actual hours worked. This means that establishments can save on labour costs when there is a short-term increase in demand, while for workers, WTAs act as an insurance against lower income during a short-term economic downturn. The use of WTAs in Germany is widespread, although it is not clear to what extent

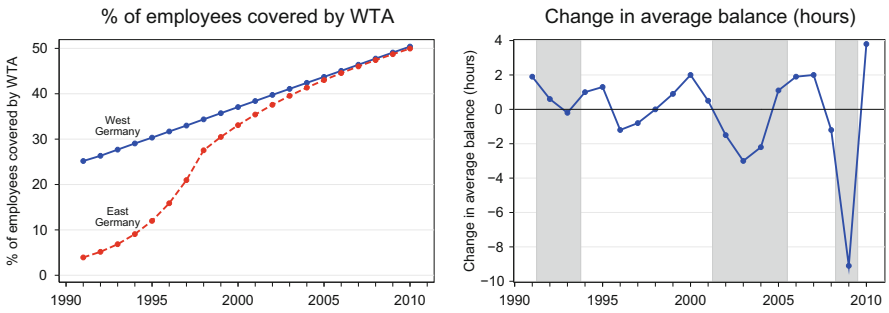


Fig. 4 Use of and average balance on Working Time Accounts 1991–2010. *Source*: Arbeitszeitrechnung des IAB, Bundesagentur für Arbeit

these are short-term "flexitime" arrangements or longer-term accounts which allow firms to adjust to demand shocks. Recent estimates Morley et al. (2009) suggest that 50 % of establishments in Germany operate WTAs, while a survey of German works councils Bogedan et al. (2009) found that changes in WTAs were the most common cost-saving method, short of redundancies, used by German establishments in the second half of 2009.

The left-hand panel of Fig. 4[6] shows that a steadily increasing proportion of workers in Germany were covered by WTA over the last 20 years, and that, by 2010, more than half of workers in both East and West Germany had such accounts. The cyclical pattern in WTA balances is clear from the right-hand panel of Fig. 4, and it is striking that the reduction in hours in the 2009 downturn was much greater than in previous downturns. Burda and Hunt (2011, Figure 10) confirm that surpluses in WTAs grew particularly quickly during the 2005–2008 expansion, and fell sharply during the 2008–2009 recession. Burda and Hunt (2011) argue that surplus WTAs provide an incentive to reduce layoffs in the short-run, because laid-off workers with a surplus of hours must be compensated. Firms therefore had an incentive to reduce the surplus before laying off workers. They also note that firms with large WTA surpluses would be less likely to use STW.

Econometric evidence on the effect of WTAs on employment adjustment is much less common than evidence on STW. One exception is Bellmann and Gerner (2011b), who compare employment growth during the crisis between plants with WTAs and those without. They find no evidence that plants which use WTAs have smaller employment adjustment, although they do find evidence of an effect in smoothing earnings, as would be expected. Boeri and Bruecker (2011) also include a measure of WTA in their plant-level employment response equations, and conclude

[6]We thank Ines Zapf for providing this information.

that WTAs also played some role in saving jobs during the great recession, although a far smaller effect than the use of STW.[7]

2.3 Bargaining Arrangements and Wage Moderation

A number of authors have stressed the role of bargaining arrangements which have allowed companies to deviate from industry-level agreements. In the past, industry-wide agreements on wages and hours of work were regarded as a source of inflexibility which prevented firms from adjusting to demand shocks. More recently, so-called "opening clauses" have been used which allowed firms and worker representatives to agree on deviations from industry-wide agreements. Typically, these take the form of "pacts for employment and competitiveness" (PEC) (Sisson and Martin Artiles 2000) whereby firms and workers agree to concessions such as reduced wages or increased hours in return for employment guarantees. Ellguth and Kohaut (2008) estimate that about 2 % of establishments in Germany have introduced a PEC, but these establishments employ around 14 % of the workforce, so they are a potentially important instrument.

Hübler (2005) and Bellmann et al. (2008) both analysed the impact of PECs on employment, while Bellmann and Gerner (2012) considered the role of PECs during the recent crisis, and found that the plants which did not exhibit employment losses during the crisis were those more likely to have adopted a PEC. However, as with STW, the selection effect is important, because establishments which enter PECs typically do so because they face serious negative demand shocks, and therefore the challenge is to find a suitable comparison group to infer what would have happened to employment in the absence of PECs.

Although there is no definitive micro-level evidence on the role of PECs, a number of authors have suggested that the wage moderation and flexibility of working time which arose during the earlier period was an important factor which allowed firms to enter the crisis without serious employment losses. Boysen-Hogrefe and Groll (2010, Figure 6), for example, shows that the real gross hourly wage increased very little from 2000 onwards. Burda and Hunt (2011) also argue that wage moderation may have played some role, though not to the same extent as Boysen-Hogrefe and Groll (2010).

2.4 Other Explanations

It seems unlikely that any of the labour market features above (STW, WTA, PECs) can, on their own, account for the resilience of the German labour market.

[7]The estimated coefficients have rather large standard errors, and therefore it is difficult to be precise about the size of the effect.

Other factors also play a role, possibly interacting with the policies described above.

For example, Klinger et al. (2011) test the hypothesis that labour hoarding was particularly common in German firms that had faced recruitment problems in earlier years. However, comparing firms which experienced labour shortages in 2008 and those that did not, Klinger et al. (2011) find no significant difference in labour hoarding behaviour during the crisis.

Möller (2010) notes that exporting firms in manufacturing industries were disproportionately affected by the downturn as a result of the collapse in export markets, and these firms tend to employ highly-skilled workers with firm-specific skills.[8] The combination of earlier recruitment problems and the potential loss of firm-specific skills increase the incentive to adopt labour hoarding. A closely-related argument in Burda and Hunt (2011) is that German manufacturing firms did not lay-off many workers in the downturn because they had hired "too few" workers in the preceding upturn of 2005–2008, probably because of weak expectations. They suggest that the weak employment increase in the pre-crisis period accounts for over a third of the difference in the employment response compared to earlier recessions.

Expectations about the length of the crisis also play a role. The crisis in Germany mainly manifested itself in terms of a decline in external demand, rather than the bursting of an asset bubble. Bohachova et al. (2011) suggest that firms expected that the shock to export demand would be short-lived, which reinforces the incentive to adopt temporary labour hoarding strategies rather than lay-off workers permanently.

Bohachova et al. (2011) use the IAB establishment panel to estimate a dynamic labour demand function for the period leading up to the crisis (2000–2008).[9] The residuals from this regression for the first half of 2009 are taken as a measure of labour hoarding, and these residuals are related to potential factors such as the use of STW and other labour market institutions. However, Bohachova et al. (2011) are not able to indicate which factors were "causally responsible" for stabilizing employment, since there is no clear counterfactual. They suggest that German firms have used multiple channels of adjustment which have been enabled by the earlier labour market reforms.

3 The Data

Our main source of data is the *Institut für Arbeitsmarkt- und Berufsforschung (IAB) Establishment Panel*. This is an annual survey of between approximately 4,000 and 10,000 establishments located in West Germany (since 1993) and between

[8]The prevalence of firm-specific vocational training in Germany may also be a factor here.

[9]The analysis is only conducted for the state of Baden-Würtemberg.

4,000 and 6,000 located in East Germany (since 1996). The sampling frame comprises all establishments in Germany with at least one worker subject to social security as of 30 June in the year before the survey. The survey currently covers approximately 1 % of all plants in Germany and approximately 7 % of workers because it is weighted towards larger plants.[10] Information is obtained by personal interviews with plant managers, and comprises about 80 questions per year, giving us information on, for example, total employment, bargaining arrangements, total sales, exports, investment, wage bill, location, and industry. In certain years specific questions are also asked about various institutional features (such as use of short-time work) and establishments' experience of the crisis.

The IAB panel also provides a measure of the total number of workers who were recruited and who left the establishment in the first half of each calendar year. In some years, information is also available on the type of workers recruited in terms of their skill level and whether they are hired on fixed-term contracts. An important advantage of the information on separations in this data is that respondents are also asked for the *cause* of the separation.[11]

We use the longest run of data available to us, from 1993 to 2010. In total, 51,603 establishments (218,570 establishment-years) appear in the survey between 1993 and 2010. We restrict the sample to those establishments in the private sector, which leaves 41,032 establishments (165,999 establishment-years).[12] We also drop establishments with missing information on employment, hires and separations, which leaves a final usable sample of 40,761 establishments (164,046 establishment-years).[13]

The relatively long run of data presents various sample selection issues. Very few establishments are followed for the entire sample period, either because of genuine establishment entry and exit, or because of sample entry and exit. In particular, the number of establishments surveyed increases substantially over time, partly as a result of the introduction of establishments in East Germany in 1996. The average size of establishment also changes over the sample period. It is therefore important to use sample weights, and to focus on within-establishment changes which control for any changes in sample composition.

[10]Weights to ensure that the sample is representative are calculated by comparing the sample of establishments with the population of establishments in the same Federal state, size and industry cell. The population of plants is obtained from a Federal Agency for Employment establishment database. A more detailed description of the data and the weighting procedure is described in Fischer et al. (2009).

[11]This includes "Dismissal on the part of the employer", "Leaving after termination of the in-company training" and "Expiration of a temporary employment contract", all of which might be regarded as dismissals by the employer. Appendix 2 gives a precise description of the relevant questions.

[12]Establishments are excluded if any of the following are true: (1) their industry is coded as "public services"; (2) profit status is coded as "non-profit"; (3) legal status is coded as "Public corporation"; (4) ownership status is coded as "Public".

[13]The sample selection procedure used is identical to that used in Bellmann et al. (2011).

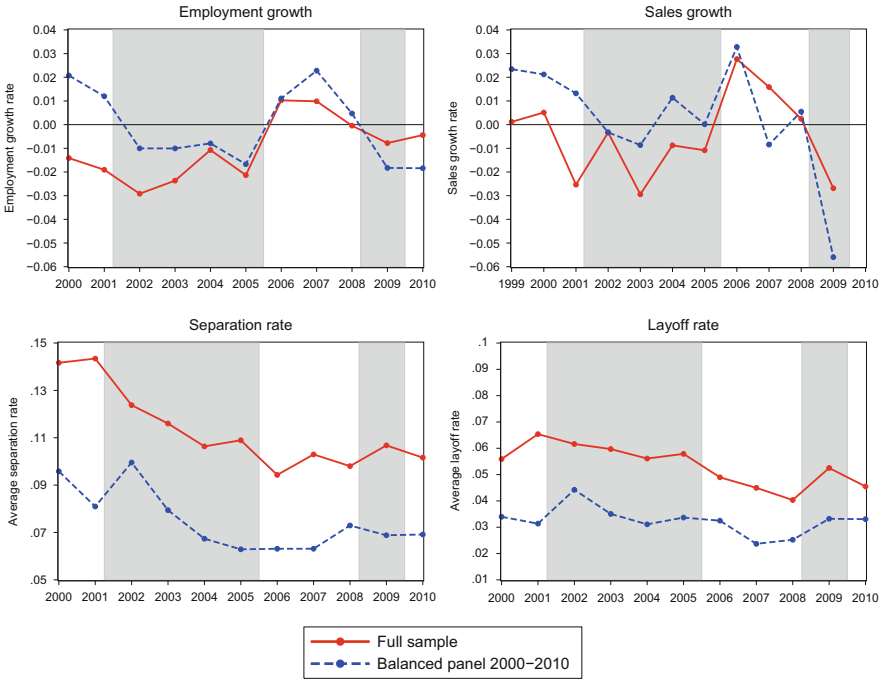


Fig. 5 Employment and sales growth rates as measured by the IAB establishment panel. Employment and sales growth are defined as $(x_{it} - x_{it-1})/0.5(x_{it} + x_{it-1})$. The full sample is weighted by cross-section weights; the balanced panel is weighted by longitudinal weights. Sales are reported for the previous calendar year and hence the series runs up to 2009. Employment refers to employment on 30 June in the current calendar year. Separations and layoffs refer to the first 6 months of the current calendar year

The top panel of Table 11 in Appendix 1 reports the number of establishments in our base sample and illustrates how the average size of establishments has declined dramatically over the sample period. The use of cross-section weights removes this fall in employment. The bottom panel of Table 11 reports the number of establishments in a balanced panel defined over the period 2000–2010. Although the use of a balanced panel greatly reduces the sample size, it is essential to be able to compare the same set of establishments before and after the crisis. The survey also includes a set of longitudinal weights which are intended to make the sample representative of the population of surviving establishments over this period.

In Fig. 5 we plot employment and sales growth rates for the last 10 years, estimated from the IAB establishment panel. Both the full sample and the balanced panel tell a similar story, although the fall in sales in the balanced panel in the most recent crisis is more severe. Nevertheless, it does appear that our sample captures the key feature of the downturn, namely a disproportionately small fall in employment relative to the fall in output.

The small fall in employment is reflected in the very weak relationship with separations and layoffs. As is well known, total separations may not be counter-cyclical because voluntary separations fall as the labour market weakens. The data show only a very small increase in layoffs between 2008 and 2009.

4 Identifying "Crisis" Establishments

The severity of the crisis for individual establishments can be captured using a self-reported measure from the survey, which asks "Have you been affected negatively by the crisis in the last 2 years?" (asked in 2010). As we would expect, this measure is highly correlated with changes in establishment sales between 2008 and 2009. Those plants which report that they were affected negatively by the crisis experienced on average a 14 % fall in sales, compared to a 1.9 % increase in sales for plants which did not report being affected.

We now examine whether establishments which were affected by the crisis differed significantly from those which were not. We do this in order to test whether the claims made in Sect. 2 about the crisis can be verified in the establishment survey. Table 1 reports the differences between crisis and non-crisis plants according to the self-reported indicator.

Crisis plants are significantly more likely to be in the manufacturing sector and significantly less likely to be in the service sector. They are significantly larger (in terms of employment and sales) and export a greater share of their output than non-crisis establishments. Crisis establishments are also more "high-tech" than non-crisis establishments, with greater R&D activity and a higher technological standard.[14]

Table 1 also gives some indication of the geographical dimension of the crisis. The two states where establishments were significantly more likely to report being affected by the crisis are Nordrhein-Westfalen and Baden-Würtemberg, both states with a high concentration of manufacturing and exporting establishments. In contrast, in Hessen, the centre of the German financial industry, establishments were less likely to report being affected by the crisis.

In Fig. 6 we plot the evolution of employment growth and sales growth for establishments separated according to their response to the crisis question. Several points are worth noting here. First of all, the two series are very similar up until 2009. This suggests that crisis establishments were not so different in terms of performance before the 2009 crisis. In particular, there was no difference in the previous downturn of 2001–2005. This result is perhaps surprising in the light of the fact that crisis establishments had quite different characteristics in terms of sector,

[14]This is based on the answer to the question "How do you assess to overall technical state of the plant and machinery, furniture and office equipment of this establishment compared to other establishments in the same industry?"

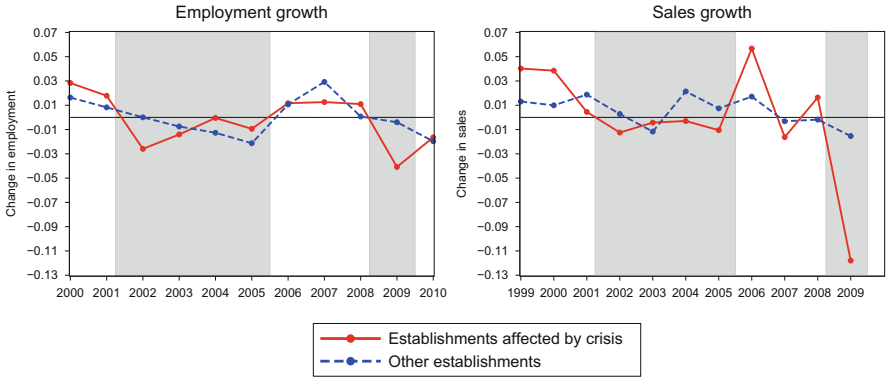


Fig. 6 Employment and sales growth separately by establishments which reported being affected by the 2008–2009 crisis and those which did not. Employment and sales growth are defined as $(x_{it} - x_{it-1})/0.5(x_{it} + x_{it-1})$. Sample used is a balanced panel of plants from 2000 to 2010, weighted by longitudinal weights. Sales are reported for the previous calendar year and hence the series run from 1999 to 2009. Employment refers to employment on 30 June in the current calendar year

size and skill intensity (see Table 1). Second, the collapse in sales is very clear in the right-hand panel, with a much smaller fall in employment, implying a large fall in labour productivity.

The argument has been made that the current crisis was "different" to earlier downturns. We have already seen that up until 2009, crisis and non-crisis plants were experiencing quite similar patterns of employment and output growth, even though crisis plants were located in different industries. Another way of seeing whether this crisis is different is to compare it with establishments' previous experience of downturns. If the 2009 crisis affected "good" establishments, we would expect that they were no more (or even less) likely to be affected by earlier downturns. We can do this by examining responses to a question about business expectations question during the period 2001–2005, shown in Table 2.[15]

Table 2 generally supports the hypothesis that establishments which were hit by the 2009 crisis were no more seriously affected by the earlier downturn. For example, 2009 crisis establishments were slightly less likely to report low business expectations for 2001 than non-crisis establishments. In four out of five cases the difference is insignificantly different from zero.

[15]This question asks: "Is business volume expected to decrease in the current year compared to the previous year?"

Table 1 Differences between crisis and non-crisis establishments

	Crisis	Non-crisis	Difference	*p*-value
Primary industries	0.0526	0.0345	0.0181	0.0557
Manufacturing	0.2300	0.1226	0.1074	0.0000
Construction	0.0959	0.1537	−0.0578	0.0003
Wholesale and retail trade	0.2324	0.2349	−0.0025	0.9019
Transport and communication	0.0775	0.0445	0.0330	0.0028
Financial and business services	0.1639	0.1791	−0.0152	0.3963
Other services	0.1477	0.2307	−0.0830	0.0000
East Germany	0.1896	0.1956	−0.0061	0.7454
Number of employees in 2007	26.3206	13.6022	12.7183	0.0005
Sales in 2007 (€M)	5.7	2.3	3.4	0.0133
Sales per worker in 2007 (€M)	0.13	0.12	0.013	0.1717
Standard hours in 2007	39.4769	39.3512	0.1257	0.1887
Net profit in 2007	0.7675	0.7497	0.0179	0.3873
Net loss in 2007	0.1658	0.1858	−0.0199	0.8615
% of sales overseas in 2007	4.6417	1.7081	2.9337	0.0000
High R& D activity	0.0692	0.0229	0.0463	0.0000
High technology standard	2.3428	2.2201	0.1227	0.0005
Share of workers with no degree	0.3596	0.3816	−0.0220	0.0448
Share of workers with university degree	0.0563	0.0394	0.0169	0.0102
Share of workers with advanced training	0.5841	0.5791	0.0050	0.6556
Independent firm	0.8322	0.9154	−0.0832	0.0000
Dependent affiliate	0.0564	0.0314	0.0250	0.0079
Firm headquarters	0.1114	0.0533	0.0581	0.0000
Schleswig-Holstein	0.0228	0.0300	−0.0072	0.3530
Hamburg	0.0300	0.0327	−0.0027	0.7489
Niedersachsen	0.0782	0.0793	−0.0011	0.9298
Bremen	0.0091	0.0038	0.0053	0.1429
Nordrhein-Westfalen	0.2613	0.1889	0.0724	0.0002
Hessen	0.0424	0.1249	−0.0825	0.0000
Rheinland-Pfalz/Saarld.	0.0554	0.0626	−0.0071	0.5255
Baden-Württemberg	0.1454	0.1072	0.0383	0.0138
Bayern	0.1657	0.1750	−0.0093	0.6034
Berlin (East and West)	0.0305	0.0345	−0.0039	0.6437
Brandenburg	0.0269	0.0301	−0.0032	0.6868
Mecklenburg-Vorpommern	0.0177	0.0149	0.0028	0.6341
Sachsen	0.0617	0.0716	−0.0099	0.4074
Sachsen-Anhalt	0.0277	0.0208	0.0069	0.3397
Thüringen	0.0249	0.0237	0.0012	0.8667

Crisis establishments are identified by the self-reported crisis indicator: "Have you been affected negatively by the crisis in the last 2 years?", asked in 2010. Characteristics refer to those recorded in 2007. Sample used is a balanced panel of 2,002 establishments observed in every year from 2000 to 2010. Weighted by longitudinal weights

Table 2 Differences between crisis and non-crisis establishments in earlier downturns

	Crisis	Non-crisis	Difference	*p*-value
Probability of reporting low business expectations for				
2001 relative to 2000	0.1467	0.1778	−0.0311	0.0789
2002 relative to 2001	0.2701	0.1758	0.0944	0.0000
2003 relative to 2002	0.1693	0.1449	0.0243	0.1576
2004 relative to 2003	0.1525	0.1801	−0.0275	0.1224
2005 relative to 2004	0.1472	0.1368	0.0104	0.5280

Crisis establishments are identified by the self-reported crisis indicator: "Have you been affected negatively by the crisis in the last 2 years?", asked in 2010. Balanced panel 2000–2010, weighted by longitudinal weights

5 Demand Shocks, Job Flows and Worker Flows

Before considering the role of various policy measures and establishment characteristics in determining the employment response to the crisis, we first describe the relationship between demand shocks, employment growth, hires and layoffs in the IAB establishment survey. This is based largely on Davis et al. (2011) and Bellmann et al. (2011). Davis et al. (2011) show that, given the relationship between employment growth, hires and separations at the establishment level, movements in the distribution of establishment employment growth rates largely drive aggregate hiring and separation rates.

5.1 The Relationship Between Job Flows and Worker Flows

For establishments in the US, falls in employment are achieved largely through an increase in separations, while increases in employment are achieved through an increase in hirings (Davis et al. 2011). Bellmann et al. (2011) show that the pattern is similar in German establishments. We replicate this basic finding for German establishments below.

The IAB panel provides a measure of the number of workers who were recruited and who left the establishment in the first 6 months of each calendar year. Define N_{it} to be employment of establishment i at time t. The net job flow, or employment change of establishment i, between $t-1$ and t, is ΔN_{it}. Employment change within an establishment will almost certainly be an underestimate of worker flows, because even for a given set of jobs, there may be workers joining and leaving the establishment. Let H_{it} (hires) be the number of workers who join the establishment between $t-1$ and t, and S_{it} (separations) be the number of workers who leave the establishment. It follows that net worker flows are equal to net job flows, $\Delta N_{it} = H_{it} - S_{it}$, but gross worker flows $H_{it} + S_{it}$ may be much larger.

It is standard to calculate separation and hiring rates by dividing by average employment between t and $t-1$:

$$h_{it} = \frac{H_{it}}{0.5(N_{it} + N_{i,t-1})}$$

$$s_{it} = \frac{S_{it}}{0.5(N_{it} + N_{i,t-1})}$$

Recall that H and S are observed over a 6-month period, and so to be consistent $t-1$ should refer to 6 months before the survey date. To ensure consistency, we define $N_{i,t-1} = N_{it} - H_{it} + S_{it}$. The net job flow rate (which equals the net worker flow rate) is then $\Delta n_{it} = h_{it} - s_{it}$. The gross worker flow rate is $h_{it} + s_{it}$ which will be greater than the net job flow rate by the amount of churning.

In Fig. 7 we plot the within-establishment relationship between employment growth (net job flows) and hiring and separation rates for the entire sample period 1993–2010. To do this we regress, separately, hiring and separation rates on a set of dummy variables for establishment growth rate bands defined by separating Δn_{it} into 50 quantiles.

The almost linear relationship between worker flows and job flows illustrated in Fig. 7 suggest the following linear spline approximation:

$$h_{it} = \alpha^h + \beta^h(\Delta n_{it} \cdot 1(\Delta n_{it} > 0)) + \gamma^h(\Delta n_{it} \cdot 1(\Delta n_{it} < 0)) + a_i^h + b_t^h + \epsilon_{it}^h \tag{1}$$

$$s_{it} = \alpha^s + \beta^s(\Delta n_{it} \cdot 1(\Delta n_{it} > 0)) + \gamma^s(\Delta n_{it} \cdot 1(\Delta n_{it} < 0)) + a_i^s + b_t^s + \epsilon_{it}^s, \tag{2}$$

where $1(\cdot)$ is the indicator function. β^h measures the responsiveness of hirings with respect to employment growth; γ^h measures the responsiveness of hirings with respect to employment falls. β^s and γ^s measure the same response with respect to separations. Because $\Delta n_{it} = h_{it} - s_{it}$ it is unnecessary to estimate both the hiring and separation equation, since $\beta^h - \beta^s = 1$ and $\gamma^h - \gamma^s = 1$. The constant in this model ($\alpha^h = \alpha^s$) is an estimate of the hiring rate (= separation rate) when establishment employment is stable over a 6-month period. Both models include establishment and time fixed-effects, a_i and b_t which can either be estimated or removed by demeaning in the usual way. The inclusion of establishment fixed effects means that the estimates of β and γ are based on within-establishment changes in job- and worker-turnover rates. We plot the estimates of the six parameters from the two linear splines given in (1) and (2) in Fig. 7.

Both the parametric and non-parametric results show a relationship which is very similar to that shown by Davis et al. (2011) for the US. We estimate β^h to be 0.98 and γ^s to be -0.96. This shows that the lack of layoffs in the crisis is *not* the result of establishments adjusting on the hiring margin. When German establishments shrink,

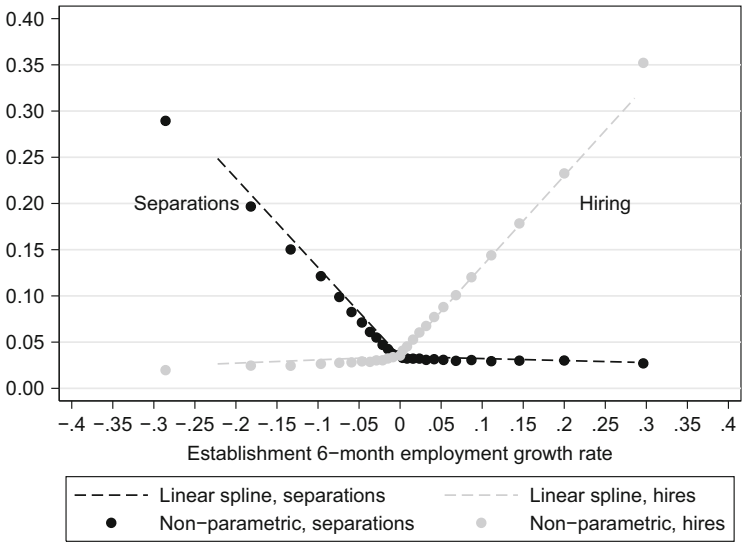


Fig. 7 Relationship between employment growth rates in the first 6 months of t with worker flow rates over the same period. Sample is 40,757 establishments and 164,019 observations. Estimated from a within-establishment fixed-effect regression with bins for each quantile of employment growth

layoffs do increase. This strong, almost symmetric relationship between worker flows and job flows holds even during the 2008–2009 period.[16]

5.2 Output Shocks and Job Flows

We therefore need to consider the relationship between the output shock and employment growth itself. We use the same methodology as above. We regress the employment growth rate (as defined above) on a set of dummy variables for sales growth rate defined by separating Δy_{it} into 50 quantiles. We also estimate the linear spline

$$\Delta n_{it} = \alpha^n + \beta^n(\Delta y_{it} \cdot 1(\Delta y_{it} > 0)) + \gamma^n(\Delta y_{it} \cdot 1(\Delta y_{it} < 0)) + a_i^n + b_t^n + \epsilon_{it}^n. \quad (3)$$

Estimates from the non-parametric and linear spline models are plotted in Fig. 8. The linear spline seems to capture the non-parametric relationship quite closely.

A number of features are worth noting. First, the relationship is quite weak. β^n is estimated to be 0.04 (0.003) and γ^n to be only 0.06 (0.003). This is partly

[16]During this period we estimate $\beta^h = 0.97$ and $\gamma^s = -0.96$.

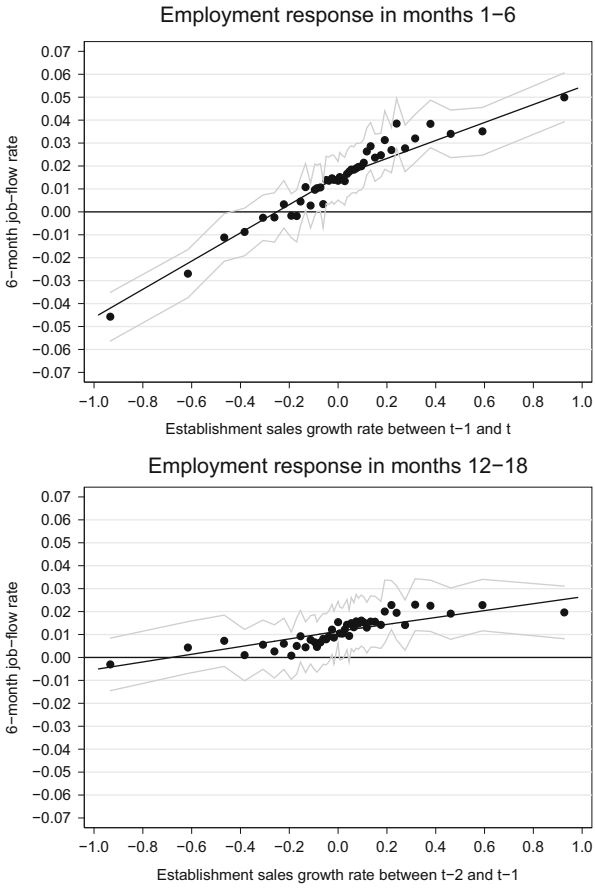


Fig. 8 Relationship between sales growth rates between $t-1$ and t and employment growth rates over the first 6 months of each year (*top panel*) and the first 6 months of the following year (*bottom panel*). Sample is 23,262 establishments and 96,353 establishment-years. Estimated from a within-establishment fixed-effect regression with bins for each quantile of sales growth which contain approximately equal numbers of observations and a linear spline

because we are looking at a very short-run relationship. The sales growth rate Δy_{it} is measured as the proportionate change in sales between years $t-1$ and t, while the job flow rate refers to the rate in the first 6 months of year t. A problem is that employment is measured as a point-in-time stock at the interview date (usually the end of June), while sales are recorded as a flow over the calendar year. Nevertheless, we can use the fact that employment at the beginning of each calendar year is equal to employment at the interview date less (hires − separations) over the preceding 6 months. Re-estimating (3) using the annual change in employment increases β^n to 0.10 (0.005) and γ^n to 0.12 (0.005). The fact that γ^n is so much smaller than one, even over the full 12 months is evidence for a strong degree of labour hoarding in the short-run. The model predicts that the stronger the negative shock to sales, the larger will be the fall in labour productivity and hence the greater the degree of labour hoarding. Note that "labour hoarding" in this context may include falls in hours per worker as well as falls in output per hour.

The second key point is that employment losses do not occur at all for small falls in sales. In fact, both the parametric and non-parametric estimates reported in Fig. 8 indicate that only quite large negative output shocks to sales result in employment losses. However, employment growth is not symmetric around $\Delta y_{it} = 0$. The response to a negative sales shock is larger than the response to a similar-size positive sales shock: in other words $\gamma^n > \beta^n$. Establishments only reduce employment in response to quite large output shocks, but they shrink faster in response to negative output shocks than they grow in response to positive output shocks.

As noted, Fig. 8 shows the very short-run relationship between employment changes in the first 6 months of year t and output changes between years $t-1$ and t. In the bottom panel of Fig. 8 we plot the relationship between employment changes in the first 6 months of year t and output changes between years $t-2$ and $t-1$, to see if lagged responses are important. It is clear that the slope is greatly reduced (β^n and γ^n are both estimated to be about 0.015, compared to be 0.05 for the contemporaneous response), so it is not the case that a large fraction of employment adjustment occurs in the year following the output shock.

The short-run relationship between output shocks (changes in sales) and employment growth provides a framework for analysing labour hoarding. When $\gamma^n < 1$ this indicates either that establishments are responding on the intensive margin (reducing hours of work per worker), or that labour productivity is falling.[17]

In Table 3 we summarise estimates of β^n and γ^n from Eq. (3). The first row reports our basic estimate of the 6-month job flow response. As noted, the response is larger for negative output shocks ($\hat{\gamma}^n > \hat{\beta}^n$). The second row shows that the 12-month job flow rates are slightly more than double the 6-month rates, although the positive (hiring) response increases more than the negative (separation) response. If

Table 3 Estimates of the short-run relationship between output shocks and job flows from Eq. (3)

	$\hat{\beta}^n$	$\hat{\gamma}^n$	N	N^*
(1) 6-month job flow rate	0.039 (0.003)	0.062 (0.003)	23,262	96,353
(2) 12-month job flow rate	0.101 (0.005)	0.123 (0.005)	23,261	96,352
(3) Additional controls	0.047 (0.004)	0.076 (0.004)	15,408	55,775
(4) 12-month job flow rate with additional controls	0.110 (0.007)	0.148 (0.007)	15,408	55,775
(5) 2008–2009	0.042 (0.010)	0.051 (0.009)	8,737	14,567

All estimates are significant at the 1% level.

[17]Note that the survey does not allow us to measure total hours of work, so we cannot distinguish changes in hours from changes in output per hour.

labour hoarding were an important phenomenon in the short-run (i.e. within-year), we would expect that $\gamma^n = 0$ over a 6-month period, but $\gamma^n > 0$ over a 12-month period. In fact, it appears that reductions in employment in response to negative output shocks over the first 6 months are slightly less than half the total response over the entire year.

In the third row and fourth rows we examine whether the relationship between job flows and output shocks is affected by the inclusion of a number of control variables.[18] As expected, the inclusion of these controls makes little difference because we are already controlling for firm fixed-effects.

In the fifth row we test whether there is any change in the response during the 2008–2009 downturn. In fact, it appears that the response is quite stable over this period. The response to negative output shocks is slightly smaller, but it is not significantly different to the response for earlier periods. Thus, in the sample as a whole there is no dramatic increase in labour hoarding during the most recent downturn.

To summarise, we find evidence that German establishments have quite a weak relationship between sales shocks and employment growth, evidence either that they adjust on the intensive margin (hours) or that they operate labour hoarding policies which entail temporary falls in labour productivity. However, we find no evidence to support the idea that establishments' behaviour was different in the 2008–2009 period. In other words, labour hoarding did not increase during this period.

5.3 Output Shocks and Worker Flows

Thus far, we have considered the relationship between output shocks and job flows (Fig. 8), and the relationship between job flows and worker flows (Fig. 7). Putting these together enables us to describe the relationship between output shocks and worker flows. In other words, how do an establishment's hiring and separations respond to output shocks? Using the same non-parametric methods as described above, Fig. 9 illustrates how hirings and separations change in response to output shocks.

As was clear from Fig. 8, in the short-run establishments continue to expand employment even when sales are static, and this is reflected in the fact that hires are about 1.5 % points higher than separations when the sales growth rate is zero. The "kink" in the hiring and separation functions is no longer as clear as in Fig. 7.

The IAB survey data also allow us to identify, to some extent, whether separations are initiated by the employer or the employee, because respondents are also asked for the cause of the separation. Appendix 2 gives a precise description of the

[18]Controls included are the change in output lagged 1 year, self-reported profitability, self-reported state of equipment in the establishment, proportion of different worker types, whether the establishment is an independent firm, bargaining arrangements and existence of a works council.

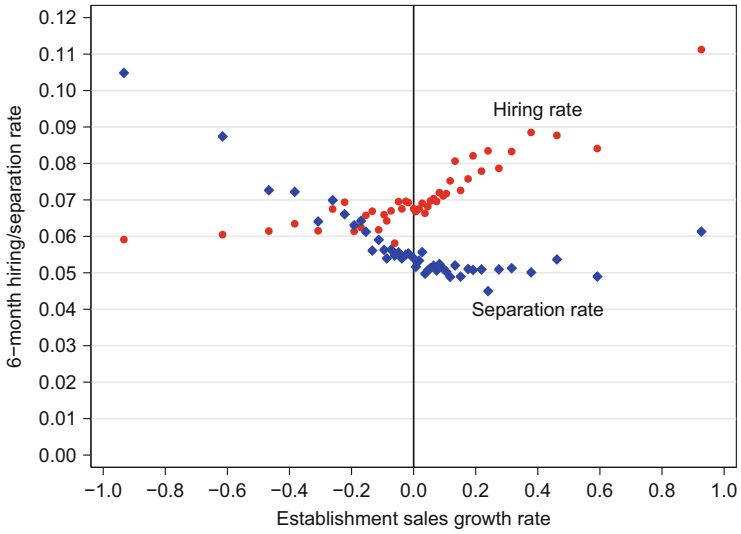


Fig. 9 Relationship between sales growth rates between $t-1$ and t and worker flows (hires and separations) over the first 6 months of each year. Sample is 23,262 establishments and 96,353 establishment-years. Estimated from a within-establishment fixed-effect regression with bins for each quantile of sales growth which contain approximately equal numbers of observations

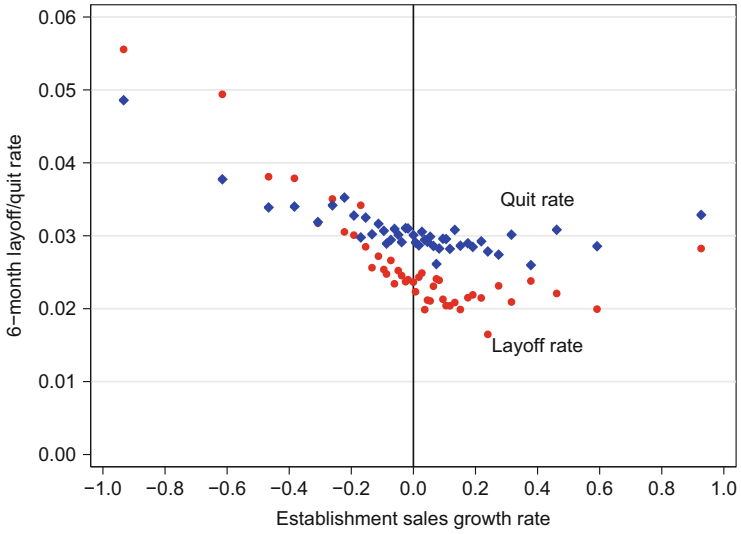


Fig. 10 Relationship between sales growth rates between $t-1$ and t, layoff and quit rates over the first 6 months of each year. Sample is 23,262 establishments and 96,353 establishment-years. Estimated from a within-establishment fixed-effect regression with bins for each quantile of sales growth which contain approximately equal numbers of observations

relevant questions. We define responses 2 (Dismissal on the part of the employer), 3 (Leaving after termination of training) and 4 (Expiration of temporary employment contract) to be layoffs, and define the remaining separations as quits.

Figure 10 shows that layoffs increase more steeply with negative sales shocks, although quits also increase to some extent. This suggests that, even in Germany, employment adjustment is not achieved solely through voluntary redundancy and a reduction in hires, but an actual in increase in layoffs.

5.4 Firm Exit

One caveat to the results presented thus far is that they rely on a sample of surviving establishments. If an establishment exits (perhaps as a result of the crisis) then they will not be interviewed in the 2010 wave, and they will not form part of the sample used to infer the effectiveness of policy measures. Furthermore, for these establishments we do not know their sales in 2009—recall that sales are reported for the previous calendar year. However, it is worth noting that the IAB establishment panel does contain a measure of exit which is potentially more reliable than that typically used in administrative data, which relies on assuming an establishment has exited if its id number no longer appears in the data. In the IAB survey, an establishment which stops responding is explicitly coded with a variable indicating whether it is no longer in business.

How do exits affect the job-flow rate? In Fig. 11 we show that the inclusion of establishment exit creates an important discontinuity in the job-flow rate. As noted,

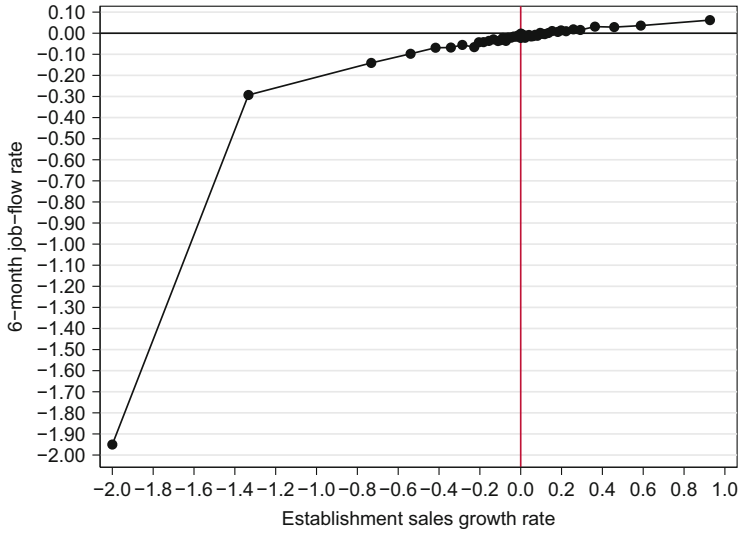


Fig. 11 Relationship between sales growth rates between $t - 1$ and t and employment growth rates over the first 6 months of each year, including establishment exit

job losses do not increase proportionately with sales shocks, but by definition the job flow rate and the sales growth rate are equal to −2 when an establishment exits.

6 The Use and Effectiveness of Policy Measures in Response to the Crisis

6.1 Use of Policy Measures

The second aim of this paper is to evaluate whether the response of establishments to the crisis differed systematically according to their use of various policy measures. An important issue here is that the use of these policy measures is not randomly assigned across establishments, and so one cannot use comparisons of establishments with and without a particular policy to infer the causal effect. We focus on several of the factors discussed in Sect. 2:

1. The use of short-time working schemes. In 2009 and 2010 the IAB establishment survey asks "Did you use Kurzarbeit in the first half of this year? If yes, how many employees have been in your short-time work program?". In addition, establishments are also asked "Have there been some further training measures combined with the Kurzarbeit programme?". Information on the use of Kurzarbeit is also available in 1993–1996, 2003 and 2006.
2. The use of working-time accounts. In 2008, 2009 and 2010 the IAB establishment survey asks "Do you have working time accounts?"[19] Establishments which do have working time accounts are also asked what proportion of employees are covered, and the time period over which the surplus and deficit have to be balanced. Information on WTA is also available in 1999, 2002, 2004 and 2006.
3. Company-level pacts on employment and competitiveness have become increasingly widely used by establishments in the twenty first century (Hübler 2005; Bellmann et al. 2008; Ellguth and Kohaut 2008; Bellmann and Gerner 2011a, 2012). If establishments have so-called "opening clauses" in their bargaining arrangements with unions, they may introduce these pacts to reduce labour costs, and they may also promote greater flexibility in employment. The IAB survey tells us whether establishments had such a pact in 2008 and 2009.

The IAB survey also asks establishments a more general question (in 2010 only) about their use of measures over the previous 2 years, and whether these measures were used in the light of the economic crisis:

1. Reduced overtime or surpluses on working time accounts
2. Increased use of holidays
3. Short-time work

[19]Possible responses: "Yes", "No", "We are planning to introduce working time accounts."

4. Other reductions in working time
5. Reductions in temporary employment
6. Increased use of further training
7. Reduced hiring or delayed employment increases
8. Layoffs

In Fig. 12 we plot the proportion of establishments using STW and WTA, which also shows which years the information is available. Both the balanced panel and the full sample tell a similar story. The proportionate increase in STW is much greater during the current crisis, but there is still a sizable increase in the use of WTA.

It is striking that the proportion of establishments which report the use of STW is actually higher in 2010 than 2009. This seems at odds with the perception that German establishments stopped using STW as soon as the recession ended. However, note that the establishment survey takes place at the end of June in each year, and asks establishments whether they used STW in the first half of that year. In the left-hand panel of Fig. 13 we plot the official statistics on the numbers of establishments using STW. One can see that the peak of about 60,000 plants is maintained until the beginning of 2010. The right-hand panel shows the number

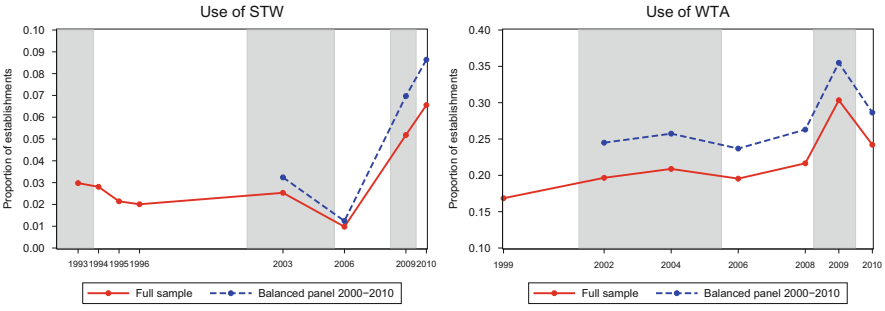


Fig. 12 Proportion of establishments using STW and WTA. The full sample is weighted by cross-section weights; the balanced panel is weighted by longitudinal weights

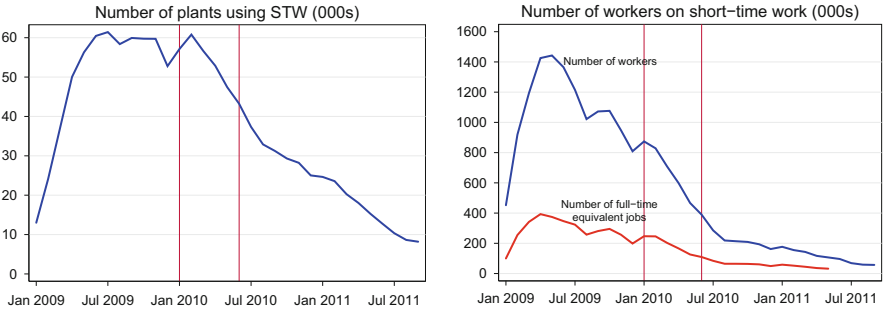


Fig. 13 Numbers of establishments and workers on STW. The *vertical lines* indicate the 6-month period to which the 2010 survey results refer. *Source*: *Bundesagentur für Arbeit*

Table 4 Differences between crisis and non-crisis establishments in the use of policy measures

	Crisis	Non-crisis	Difference
Establishment uses short-time work	0.14	0.03	0.11***
Proportion of employees covered	0.06	0.01	0.05***
Establishment uses WTA	0.39	0.33	0.06***
Proportion of employees covered	0.35	0.29	0.05**
Establishment uses PECs	0.03	0.01	0.02***
Responses to 2010 question			
Reduced overtime or surpluses on working time accounts	0.32	0.09	0.23***
Reduced hiring or delayed employment increases	0.31	0.06	0.26***
Short time work	0.24	0.03	0.22***
Increased use of holidays	0.23	0.05	0.18***
Layoffs	0.15	0.05	0.09***
Other reductions in working time	0.12	0.02	0.10***
Reductions in temporary employment	0.11	0.01	0.09***
Increased use of further training	0.08	0.03	0.05***
Layoff trainees at end of training programme	0.04	0.01	0.03***

Crisis establishments are identified by the backward-looking self-reported crisis indicator: "Have you been affected negatively by the crisis in the last 2 years?", asked in 2010. Balanced panel 2000–2010, weighted by longitudinal weights. *** and ** indicate significance at the 1% and 5% level respectively.

of workers on STW; this drops off much earlier in 2009. Thus, we would expect estimates from the establishment survey to only start dropping in the 2011 survey.

Table 4 summarises the use of these various policy measures in 2009, separately for crisis and non-crisis establishments. Unsurprisingly, crisis plants were nearly five times more likely to have used STW in 2009, although for those that did use STW, the difference in the proportion of workers covered was smaller. There is also a significant difference in the use of WTA, but this difference is much smaller, which presumably reflects the fact that WTAs were introduced as a result of negotiations with labour unions, rather than explicitly as a crisis measure.

The proportion of establishments using company-level pacts for employment and competitiveness is also significantly higher amongst crisis plants, but the overall proportion using them is small. Note that PECs were predominantly used by larger establishments, and these results are weighted to reflect the population of establishments, which comprises many more small establishments.

The bottom panel of Table 4 shows the proportion of crisis and non-crisis plants which reported having used various measures in response to the crisis. These are ordered by the frequency of use amongst crisis establishments. It is striking that the most common response for crisis plants was to reduce overtime or surpluses on working time accounts. The second most common response was to reduce hiring or delay employment increases, with the use of short-time work the third most widely-used measure. These responses confirm that only a small fraction of plants resorted to layoffs in response to the crisis.

A more systematic examination of the determinants of the use of STW and WTA is presented in Tables 5 and 6.[20] In Table 5 we report estimates of a Probit model for the probability of using STW in 2009, and estimates of a Tobit model for the proportion of workers covered by STW.[21] These results confirm that various measures of firm performance up to 2009 are important determinants of whether or not an establishment uses STW in 2009, and the proportion of workers affected. For example, the expected growth rate of turnover between 2008 and 2009 is negatively associated with use of STW, and an indicator for turnover decline between 2007 and 2008 is positively associated with use of STW. Self-reported profitability also shows that negative shocks are strongly associated with use of STW. Establishments which report "unsatisfactory" profits for 2008 are 10 % points more likely to use STW in 2009 than those reporting "very good" profits. The proportion of workers affected by STW is more than 30 % points higher for these establishments.

We also note that various measures of the "flexibility" of the existing workforce are negatively related to the use of STW. For example, establishments with a higher proportion of workers on fixed-term contracts and agency workers are less likely to adopt STW, which suggests that STW is a substitute for the flexibility which comes from having a short-term workforce.[22]

Table 5 also shows that establishments which reported labour shortages in 2008 were *not* significantly more likely to use STW, and in fact had a smaller proportion of workers on STW. We also do not find a significant effect for exporting plants or for "high-tech" plants. These insignificant results show that any relationship in the raw data is driven by the industry of establishments, which we are controlling for here.[23]

We noted earlier that there was a geographical concentration of crisis establishments, particularly in Baden-Württemberg and Nordrhein-Westfalen (see Table 1). However, it is noticeable that there are few significant state-effects on either the probability of using STW or the proportion of workers affected, suggesting that any state-level factors are being captured by the establishment-level measures in the model.

Table 6 estimates the same models for the probability of using WTA and the proportion of workers covered by WTA. Note that the coefficient estimates are quite different, indicating that different factors were associated with the use of WTA. Poorly-performing establishments are not more likely to use WTA, and in fact use of WTA and the proportion of workers covered increase rather than decrease with profitability. This confirms that WTA is not used as an emergency crisis measure, but relates to longer-term negotiations between establishments and unions. Note,

[20]This is a similar model to that estimated by Boeri and Bruecker (2011).

[21]The Tobit model interprets the proportion of workers covered as a continuous variable censored at zero.

[22]Deeke (2005) finds a similar result based on earlier waves of the IAB survey.

[23]The coefficients on industry are large and highly significant for manufacturing industries in these regressions.

Table 5 Probit estimates of the probability of using STW and Tobit estimates of the proportion of workers affected by STW in the first 6-months of 2009

	Prob. of using STW		Prop. of workers affected by STW	
	Marg. eff.	S.E.	Marg. eff.	S.E.
Expected growth rate of turnover 2008–2009	−0.003***	(0.000)	−0.015***	(0.001)
Actual turnover declined between 2007 and 2008 (1 = yes)	0.040***	(0.007)	0.154***	(0.030)
Self-reported risk of firm closure (1 = yes)	0.015	(0.010)	0.028	(0.040)
High competitive pressure 2009 (1 = yes)	0.009	(0.008)	0.064	(0.036)
High competitive pressure 2008 (1 = yes)	−0.008	(0.007)	−0.041	(0.032)
Proportion of output exported	0.000	(0.000)	0.001	(0.001)
Self-reported profitability 2008 (1 = very good)				
2 Good	0.012	(0.011)	0.051	(0.048)
3 Satisfactory	0.025*	(0.013)	0.106*	(0.052)
4 Sufficient	0.060**	(0.019)	0.235***	(0.058)
5 Unsatisfactory	0.095***	(0.028)	0.345***	(0.069)
Labour shortages reported in 2008 (1 = yes)	−0.009	(0.007)	−0.081*	(0.032)
High share of R&D activities (1 = yes)	0.016	(0.010)	0.065	(0.038)
Technical state of establishment (1 = state of the art)				
2	0.008	(0.009)	0.029	(0.042)
3	0.018	(0.011)	0.100*	(0.046)
4	0.043	(0.025)	0.170*	(0.079)
5 (obsolete)	0.019	(0.063)	0.045	(0.373)
Proportion of qualified workers	−0.031*	(0.014)	−0.145*	(0.066)
Proportion of women	−0.034*	(0.015)	−0.075	(0.077)
Proportion of part-time workers	−0.095***	(0.023)	−0.670***	(0.123)
Proportion of fixed-term workers	−0.087*	(0.044)	−0.404*	(0.196)
Proportion of agency workers	−0.114	(0.062)	−0.598*	(0.300)
Proportion of owners working in plant	−0.148***	(0.032)	−0.786***	(0.167)
Independent plant (1 = yes)	0.005	(0.011)	0.050	(0.052)
Headquarters (1 = yes)	−0.020	(0.011)	−0.043	(0.060)
Firm managed by owner (1 = yes)	0.017	(0.010)	0.059	(0.047)
Firm managed by prof. manager (1 = yes)	0.008	(0.012)	−0.014	(0.049)
Chamber of commerce membership (1 = yes)	0.011	(0.017)	0.056	(0.095)
Firm-level bargaining (1 = yes)	0.004	(0.012)	−0.031	(0.052)
Industry-level bargaining (1 = yes)	−0.004	(0.007)	−0.035	(0.036)
Works council (1 = yes)	0.001	(0.009)	0.060	(0.042)
Bundesland (1 = Schleswig-Holstein)				
Hamburg	0.018	(0.036)	0.030	(0.136)
Niedersachsen	0.039	(0.027)	0.112	(0.089)
Bremen	0.026	(0.029)	0.045	(0.115)

(continued)

Table 5 (continued)

	Prob. of using STW		Prop. of workers affected by STW	
	Marg. eff.	S.E.	Marg. eff.	S.E.
Nordrhein-Westfalen	0.012	(0.021)	−0.032	(0.088)
Hessen	0.043	(0.029)	0.125	(0.091)
Rheinland-Pfalz/Saarld.	0.018	(0.024)	0.044	(0.091)
Baden-Württemberg	0.028	(0.025)	0.044	(0.089)
Bayern	0.007	(0.022)	−0.018	(0.092)
Berlin (East and West)	−0.018	(0.019)	−0.203*	(0.120)
Brandenburg	0.016	(0.023)	0.008	(0.095)
Mecklenburg-Vorpommern	0.006	(0.024)	−0.018	(0.108)
Sachsen	0.059**	(0.029)	0.219**	(0.087)
Sachsen-Anhalt	0.015	(0.022)	−0.017	(0.093)
Thüringen	0.046*	(0.027)	0.123	(0.087)
Pseudo-R^2	0.379		0.359	
N	6,156		6,087	

Estimates also include dummies for 1-digit sector and (log) employment in 2008 and 2009. Marginal effects for the Tobit model refer to the marginal effect on the censored mean; see Cameron and Trivedi (2009, p. 527). ***, ** and * indicate significance at the 1%, 5% and 10% level respectively.

Table 6 Probit estimates of the probability of using WTA and Tobit estimates of the proportion of workers using WTA in the first 6-months of 2009

	Prob. of using WTA		Prop. of workers affected by WTA	
	Marg. eff.	S.E.	Marg. eff.	S.E.
Expected growth rate of turnover 2008–2009	−0.000	(0.000)	−0.001	(0.001)
Actual turnover declined between 2007 and 2008 (1 = yes)	0.012	(0.015)	0.020	(0.021)
Self-reported risk of firm closure (1 = yes)	0.000	(0.022)	−0.004	(0.029)
High competitive pressure 2009 (1 = yes)	0.001	(0.018)	0.004	(0.025)
High competitive pressure 2008 (1 = yes)	0.051**	(0.016)	0.069**	(0.022)
Proportion of output exported	−0.001	(0.000)	−0.001*	(0.001)
Self-reported profitability 2008 (1 = very good)				
2 Good	−0.052	(0.027)	−0.064	(0.034)
3 Satisfactory	−0.052	(0.028)	−0.082*	(0.036)
4 Sufficient	−0.073*	(0.031)	−0.112**	(0.042)
5 Unsatisfactory	−0.107**	(0.037)	−0.182***	(0.050)
Labour shortages reported in 2008 (1 = yes)	0.020	(0.016)	0.050*	(0.022)
High share of R&D activities (1 = yes)	0.091***	(0.024)	0.107***	(0.029)

(continued)

Table 6 (continued)

	Prob. of using WTA		Prop. of workers affected by WTA	
	Marg. eff.	S.E.	Marg. eff.	S.E.
Technical state of establishment (1 = state of the art)				
2	−0.010	(0.020)	−0.016	(0.027)
3	−0.025	(0.022)	−0.034	(0.031)
4	0.011	(0.042)	0.017	(0.059)
5 Obsolete	−0.291*	(0.140)	−0.490	(0.333)
Proportion of qualified workers	0.104***	(0.031)	0.143**	(0.045)
Proportion of women	−0.130***	(0.032)	−0.179***	(0.050)
Proportion of part-time workers	−0.107**	(0.038)	−0.194**	(0.063)
Proportion of fixed-term workers	0.037	(0.077)	0.097	(0.103)
Proportion of agency workers	0.107	(0.135)	0.167	(0.151)
Proportion of owners working in plant	−0.442***	(0.072)	−0.997***	(0.117)
Independent plant (1 = yes)	0.026	(0.031)	0.048	(0.037)
Headquarters (1 = yes)	0.079*	(0.037)	0.108*	(0.042)
Firm managed by owner (1 = yes)	−0.057	(0.029)	−0.059	(0.035)
Firm managed by prof. manager (1 = yes)	0.002	(0.032)	0.023	(0.035)
Chamber of commerce membership (1 = yes)	0.063	(0.033)	0.080	(0.050)
Firm-level bargaining (1 = yes)	0.101**	(0.032)	0.119***	(0.036)
Industry-level bargaining (1 = yes)	0.060***	(0.017)	0.077**	(0.024)
Works council (1 = yes)	0.066**	(0.025)	0.076*	(0.030)
Bundesland (1 = Schleswig-Holstein)				
Hamburg	−0.118*	(0.062)	−0.096	(0.089)
Niedersachsen	−0.007	(0.043)	0.009	(0.060)
Bremen	0.010	(0.045)	0.050	(0.066)
Nordrhein-Westfalen	−0.067*	(0.041)	−0.085	(0.058)
Hessen	−0.054	(0.046)	−0.035	(0.066)
Rheinland-Pfalz/Saarld.	−0.074*	(0.042)	−0.118*	(0.062)
Baden-Württemberg	0.083**	(0.041)	0.148**	(0.060)
Bayern	0.011	(0.043)	0.028	(0.060)
Berlin (East and West)	−0.084*	(0.047)	−0.114	(0.072)
Brandenburg	0.054	(0.043)	0.089	(0.062)
Mecklenburg-Vorpommern	−0.002	(0.045)	0.010	(0.067)
Sachsen	0.056	(0.042)	0.107*	(0.060)
Sachsen-Anhalt	0.031	(0.042)	0.062	(0.060)
Thüringen	−0.041	(0.042)	−0.016	(0.061)
Pseudo-R^2	0.201		0.132	
N	6,180		6,161	

Estimates also include dummies for 1-digit sector and (log) employment in 2008 and 2009. Marginal effects for the Tobit model refer to the marginal effect on the censored mean; see Cameron and Trivedi (2009, p. 527). ***, ** and * indicate significance at the 1%, 5% and 10% level respectively.

for example, that the use of WTA is positively associated with firm- and industry-level bargaining measures, the existence of a works council and establishment size in 2008. Nevertheless, WTA may still be a way of dealing with a temporary decline in labour demand without resorting to declines in employment.

6.2 The Impact of Policy Measures

6.2.1 STW

As noted in Sect. 2, an assessment of the impact of STW on the response to the crisis is plagued by an extreme selection problem. Establishments which used STW did so in part because they experienced a larger fall in sales in 2009. Figure 14 illustrates this clearly. In the left-hand panel we plot the density of sales growth between 2008 and 2009 for establishments which reported being affected by the crisis, split between those which used STW in 2009 and those that did not.

Sales growth is substantially worse for the STW establishments. Our estimates of the employment response to sales shocks in Sect. 5 suggest that larger negative shocks to sales will cause greater labour hoarding (a greater fall in labour productivity) because $\gamma^n < 1$. In the right-hand panel of Fig. 14 we show that this is exactly what happened. The STW establishments had larger falls in labour productivity (greater labour hoarding) during 2009. However, it is important to realise that this observed labour hoarding is *not* the result of the use of STW. Instead, it is a function of the greater output shock faced by these establishments. In fact, the difference in the fall in labour productivity between the two types of establishment is *smaller* than the difference in the fall in sales, which suggests that, for a given fall in sales, STW establishments reduced employment more than non-STW establishments.

To examine the effectiveness of STW on job and worker flows more formally, we use a difference-in-difference variant of Eq. (3). We first separate establishments

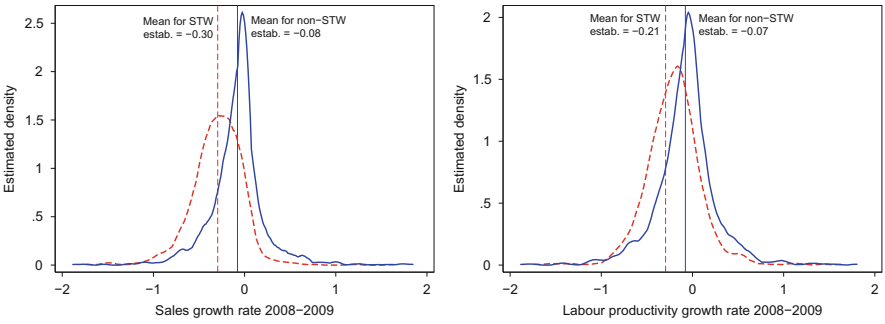


Fig. 14 Distribution of sales growth and labour productivity growth rates between 2008 and 2009 for establishments which reported being affected by the crisis, split between those which reported using STW in 2008 and those that did not

into a treatment and a control group. The treatment group are those establishments which made use of STW in the first 6 months of 2009, and the control group are those which did not. Define STW_i as a dummy variable which takes the value 1 for establishments in the treatment group and zero otherwise. Define D_t^{09} as a dummy variable which takes the value 1 in 2009 and zero otherwise. To simplify notation, define $y_{it}^{+} \equiv \Delta y_{it} \cdot 1(\Delta y_{it} > 0))$ and $y_{it}^{-} \equiv \Delta y_{it} \cdot 1(\Delta y_{it} < 0))$. Interacting Eq. (3) with the treatment indicator and the indicator for 2009 gives us the following difference-in-difference model:

$$\Delta n_{it} = \alpha_1^n + \beta_1^n y_{it}^{+} + \gamma_1^n y_{it}^{-} + \beta_2^n (y_{it}^{+} D_t^{09}) + \gamma_2^n (y_{it}^{-} D_t^{09}) + \beta_3^n (y_{it}^{+} \text{STW}_i) + \gamma_3^n (y_{it}^{-} \text{STW}_i) + \beta_4^n (y_{it}^{+} \text{STW}_i D_t^{09}) + \gamma_4^n (y_{it}^{-} \text{STW}_i D_t^{09}) + a_i^n + \epsilon_{it}^n. \tag{4}$$

The treatment indicator STW_i is not included because we include a full set of establishment fixed-effects a_i^n. In this model, the parameters we are most interested in are γ_1 to γ_4, which give the response of employment to a negative output shock in the control group and the treatment group before and during the crisis. γ_4 is the difference-in-difference estimate of this response. If establishments in the treatment group have smaller responses to output shocks because they practice greater labour hoarding, then we would predict $\gamma_4 < 0$.

The first column of Table 7 reports estimates of γ_1–γ_4 from Eq. (4) for a balanced panel of establishments observed in every year from 2005 to 2009. $\hat{\gamma}_1 = 0.038$, which is the estimated employment fall in response to an output shock for the control group in the "before" period (i.e. before 2009). γ_2 is estimated to be zero, which means that the response to output shocks did not change for the control group in 2009. Similarly γ_3 is also estimated to be zero, which means that the response to output shocks for the treatment group was similar to the control group in the pre-crisis period. But γ_4 is estimated to be positive and significant at the 5 % level, indicating that STW establishments had *larger* falls in employment for a given output shock.

The remaining columns of Table 7 decompose this job flow differential into its constituent worker flows: hires, separations and layoffs. These are essentially parametric estimates of the relationships plotted in Figs. 9 and 10. Note that the response of hires to negative sales shocks is small, and this is reflected in the small and generally insignificant estimates in the second column. Larger responses come through the separation function, and a majority of that response is due to increases in layoffs. For example, the total estimated treatment effect on job flows is 0.081. This can be decomposed into a steeper (more negative) separation response of −0.039 and a steeper (more positive) hiring response of 0.043. Thus about $0.039/0.081 = 48\,\%$ of the increased job loss of STW establishments came from increased separations, while the remaining 52 % came from reduced hires. Further, a

Table 7 Difference-in-difference estimates of the impact of STW on job flows hires and separations, estimated from Eq. (4)

	Job flows	Hires	Separations	Layoffs
$\hat{\gamma}_1$	0.038***	0.012	−0.025**	−0.017**
	(0.014)	(0.011)	(0.011)	(0.009)
$\hat{\gamma}_2$	0.004	0.006	0.002	0.004
	(0.021)	(0.016)	(0.017)	(0.012)
$\hat{\gamma}_3$	0.000	−0.021	−0.021	−0.038
	(0.038)	(0.028)	(0.035)	(0.033)
$\hat{\gamma}_4$	0.081**	0.043	−0.039	−0.021
	(0.039)	(0.030)	(0.036)	(0.032)
Estab. with <50 % coverage:				
$\hat{\gamma}_4$	0.141***	0.063	−0.079**	−0.072**
	(0.049)	(0.035)	(0.039)	(0.031)
Estab. with 50–75 % coverage:				
$\hat{\gamma}_4$	0.177***	0.092	−0.085**	−0.046
	(0.070)	(0.061)	(0.035)	(0.029)
Estab. with > 75 % coverage:				
$\hat{\gamma}_4$	−0.045	−0.019	0.026	0.041
	(0.064)	(0.048)	(0.071)	(0.068)

Sample is a balanced panel of 3,470 establishments observed over the period 2005–2009. The regressions include the full set of controls used to estimate the probability of using STW (see Table 5). ***, ** and * indicate significance at the 1%, 5% and 10% level respectively.

majority ($-0.021/-0.039 = 54\,\%$) of these separations can be attributed to layoffs rather than quits.

What can explain this finding? The rules governing the use of the STW scheme stipulate that workers may not be laid off during the period that STW is in use. However, our results suggest that STW only directly protects the jobs of those workers who are on short-time work.[24] In 2009 there were 1,202 establishments in the sample which stated that they used STW in the first 6 months of that year. Of these, 711 had job losses over that same 6 month period. But only 3 establishments had job losses which were greater than the number of non-STW workers.[25]

Most establishments only used STW for a minority of their workforce. In Fig. 15 we plot the distribution of "unprotected jobs" amongst STW establishments in 2009. 75 % of establishments had less than 25 % of their jobs protected by STW programs, and so for the great majority it was possible to continue to make employment reductions despite the use of STW. Employment reductions of more than 10 % are

[24] A second possible explanation is that the employment adjustment (which is recorded over the first 6 months of 2009) occurred before the use of STW if STW started after January of 2009.

[25] We assume that this is caused by measurement error, either in the variable recording employment or the variable recording the number of workers covered by STW.

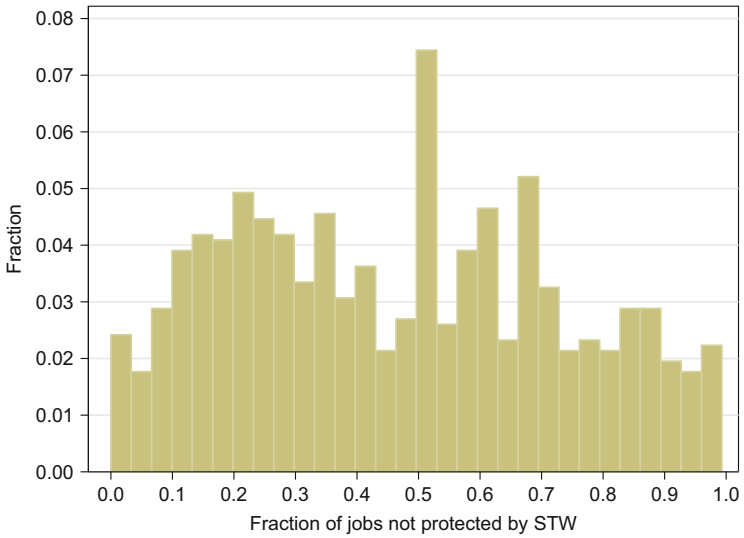


Fig. 15 Fraction of jobs which were not protected by STW programs in 2009 amongst those establishments which operated STW. 75 % of establishments have less than 25 % of their jobs protected by STW programs

unusual even in the face of large negative shocks to sales, and so it appears that STW did not prevent this kind of adjustment.

In the bottom panel of Table 7 we show that the additional employment losses of STW establishments are confined to those establishments with a smaller fraction of workers covered by STW. Splitting the treatment group into three according to the proportion of workers covered reveals that the treatment effect is large and positive for those establishments with less than three-quarters of their workforce on STW. For establishments with more than three-quarters on STW, employment losses are estimated to be smaller than in the control group ($\hat{\gamma}_4 = -0.076$), although this is imprecisely estimated and insignificantly different from zero.

However, it is clear that the additional job losses incurred by STW establishments is almost certainly the result of selection into the program. Although the estimate controls for pre-existing differences in adjustment response (by using a DiD), and is conditional on the change in sales, it still seems plausible that STW establishments faced larger (unobserved) negative shocks in addition to those captured by the change in sales. To deal with this problem requires an instrument which is correlated with the decision to use STW in 2009, but which is not correlated with the shock in 2009. Boeri and Bruecker (2011) suggest that prior use of STW may be a suitable instrument. First, prior use of STW increases indicates that establishments are familiar with the program and are more likely to use it in 2009. Second, Boeri and Bruecker (2011) argue that the 2009 shock was uncorrelated with earlier output shocks which caused prior use of STW.

Table 8 Comparison of OLS and IV difference-in-difference estimates of the impact of STW on job flows, estimated from Eq. (4)

	OLS	IV
$\hat{\gamma}_1^n$	0.038***	0.058
	(0.014)	(0.037)
$\hat{\gamma}_2^n$	0.004	−0.001
	(0.021)	(0.055)
$\hat{\gamma}_3^n$	0.000	0.076
	(0.038)	(0.248)
$\hat{\gamma}_4^n$	0.081**	−0.059
	(0.039)	(0.241)

Sample is a balanced panel of 3,470 establishments observed over the period 2005–2009. STW use in 2009 instrumented with use in 2006 and 2003. The regressions include the full set of controls used to estimate the probability of using STW (see Table 5). ***, ** and * indicate significance at the 1%, 5% and 10% level respectively.

Some evidence to support this idea is available by comparing the pattern of sales shocks for establishments which use STW in 2009 and those which used STW in 2006. As we have already seen, establishments which used STW in 2009 had much larger falls in sales in that year. However, establishments which used STW in 2006 have almost exactly the same fall in sales in 2009 as establishments which did not use STW in 2006. But use of STW in 2006 is strongly correlated with use of STW in 2009.[26]

Table 8 compares the OLS and IV estimates of the DiD model. Unfortunately, the explanatory power of the instruments in the first stage regression is not enough to draw any reliable conclusions from these results.[27] Essentially, we find that prior use of STW does not have enough explanatory power to predict use of STW in 2009. The IV estimates are therefore very imprecise, although we do find that the DiD estimate is now negative rather than positive, as expected.

[26] A regression of STW (2009) on STW (2006) has a coefficient of 0.29 with a standard error of 0.04.

[27] There are five endogenous variables in the DiD model (STW, STW interacted with y^- and y^- and STW interacted with y^+D^{09} and y^-D^{09}. The F statistics from these first stage regressions are always less than the recommended value.

6.2.2 WTA, PEC and Labour Shortages

In addition to STW, we are also interested to see if the response to sales shocks was ameliorated in establishments which used WTA, which had negotiated a pact for employment and competitiveness (PEC) and which had experience labour shortages in the period leading up to the 2008–2009 downturn. Each of these factors has been suggested as a possible cause of the apparent resilience of the German labour market. To evaluate the impact of these factors we use a simplified version of Eq. (4) which estimates β and γ separately for establishments in the treatment and control group. In the case of WTA, we have:

$$\Delta n_{it} = \alpha_1^n + \beta_1^n y_{it}^+ + \gamma_1^n y_{it}^- + \beta_2^n (y_{it}^+ \text{WTA}_i) + \gamma_2^n (y_{it}^- \text{WTA}_i) + a_i^n + D_t^{09} + \epsilon_{it}^n. \quad (5)$$

In this case it is less appropriate to use a DiD methodology because the characteristics of the establishment are already set before the 2008–2009 crisis begins. Equation (5) therefore simply compares the slope of the adjustment response between different types of plant in 2009.

In Table 9 we report estimates of γ_2 from Eq. (5). Establishments with WTA in place in 2008 have significantly larger job flows for a given fall in sales ($\hat{\gamma}_2^n = 0.035(0.019)$. Taken at face value, this contradicts the assertion that WTA helped to protect jobs. However, as with the result for STW, we suspect that greater negative

Table 9 OLS estimates of the impact of WTA, PEC and labour shortages on job and worker flows, estimated from Eq. (5)

	Job flows $\hat{\gamma}_2^n$	Hires $\hat{\gamma}_2^h$	Separations $\hat{\gamma}_2^s$	Layoffs $\hat{\gamma}_2^l$
Used WTA in 2008	0.035*	−0.001	−0.035**	−0.025**
	(0.019)	(0.016)	(0.014)	(0.010)
Used WTA in 2008:				
<80 % of workers covered	0.018	0.013	−0.005	0.004
	(0.029)	(0.023)	(0.022)	(0.016)
80–100 % of workers covered	0.045	−0.007	−0.052**	−0.039**
	(0.030)	(0.026)	(0.023)	(0.017)
All workers covered	0.041*	−0.003	−0.044**	−0.035***
	(0.023)	(0.019)	(0.017)	(0.013)
Utilised a PEC in 2008	−0.034	−0.026	0.008	0.017
	(0.035)	(0.029)	(0.026)	(0.019)
Experienced labour	0.044**	0.026*	−0.018	−0.019*
shortages in 2008	(0.019)	(0.016)	(0.014)	(0.010)

The coefficients reported are γ_2^n from Eq. (5), and thus represent the additional impact of the treatment on flows relating to negative sales shocks. The regressions include the full set of controls used to estimate the probability of using WTA (see Table 6). ***, ** and * indicate significance at the 1%, 5% and 10% level respectively.

selection effects could explain this result. We also note that, unlike STW, there is no evidence that having a greater proportion of the workforce covered by WTA helps to make employment more resilient to output shocks. The estimates of γ_2^n are actually larger for establishments with a greater proportion of the workforce covered. As with STW, the relationship between WTA and employment adjustment comes through separations and not hires. None of the estimates of γ_2^h are significantly different from zero, while establishments with WTA did have significantly more separations. Furthermore, the increase in separations is very similar to the increase in layoffs.

In contrast, establishments which had agreed a PEC in 2008 produce a negative estimate of γ_2^n, indicating that these establishments had smaller job flows. However, the estimate is imprecise. We cannot reject the hypothesis that γ_2^n is equal for PEC and non-PEC establishments, but nor can we reject the hypothesis that $\gamma_1^n + \gamma_2^n = 0$, which would indicate that establishments with a PEC had a zero short-run employment response to output shocks. These results are consistent with those found by Bellmann and Gerner (2012).

Finally, we consider whether establishments which reported labour shortages in 2008 were more likely to hoard labour in 2009 (see Sect. 2.4 for a discussion of this hypothesis.) We do not have a precise measure of labour shortages, but establishments were asked whether the company could have achieved an increase in turnover with the resources available in that year. We code an establishment as having labour shortages if it stated that it would *not* have been possible to meet an increase in output without hiring more staff. We do not find any support for the notion that establishments which faced labour shortages (or at least those that did not have excess labour) had smaller employment losses. In fact, the estimate of γ_2^n is positive and significant.

7 Policy Measures and the Pattern of Recovery

The third objective of this chapter is to examine the implications of the use of STW and WTA for the early recovery period. As noted by Hijzen and Venn (2011), one of the main concerns about the use of STW is that it may inhibit reallocation of employment and growth, if the schemes are allowed to continue for a longer period. We note that this worry does not seem particularly relevant for Germany, where the use of STW was time-limited, and where the use of STW was already declining by the start of 2010.

We can only draw tentative conclusions here, mainly because of data availability. Although the survey results for 2010 are now available at the time of writing, information on sales performance is only available for 2009. To assess whether establishments have successfully recovered from the crisis, we make use of a forward-looking question asked in each year: “How do you expect business volume to develop in the current year, as compared to the previous year?”. Respondents are

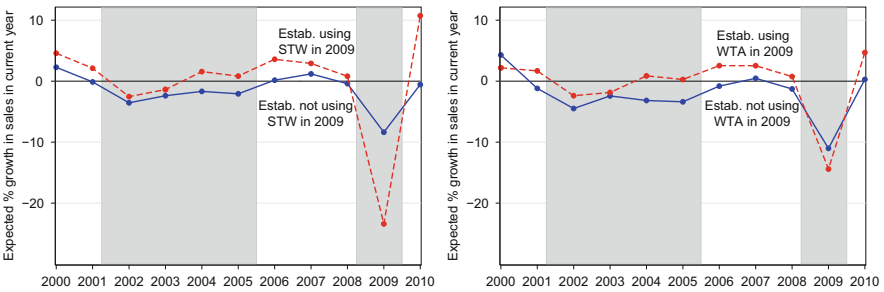


Fig. 16 Expected change in sales 2000–2010, for establishments which reported being affected by the crisis, split between use of STW in 2008 and those that did not (*left-hand panel*) and by use of WTA in 2008 (*right-hand panel*)

also asked for an estimate of the percentage change in sales between the current and the previous year.

In Fig. 16 we plot the development of these business expectations over time, including the most recent year. The left hand-panel shows how business expectations of establishments which used STW in 2009 plummeted. A comparison with the retrospective reported change in sales between 2008 and 2009 (Fig. 14) shows that these expectations were accurate, on average. Figure 16 also shows that STW establishments expected to recover quite strongly in 2010, in particular in comparison with non-STW establishments. The right-hand panel makes a similar comparison for WTA and non-WTA establishments. As noted before, WTA was much less of a crisis instrument, and therefore differences are much smaller.

A further simple test of whether establishments which used STW or WTA in 2008 faced ongoing problems is to examine the responses to the question:"What kind of problems with human resources management do you expect for your establishment/office during the next 2 years?" Responses include: whether the establishment was expected to face too high a staff level, too much staff turnover, difficulties in hiring qualified workers, and high wage costs. We estimate a linear model of the form:

$$\Pr(P_i = 1) = \alpha + \beta \mathrm{STW}_i + \gamma \mathrm{WTA}_i + \mathbf{x}_i' \boldsymbol{\delta} + \epsilon_i \tag{6}$$

where P_i is a dummy variable indicating that human resource problems are a potential problem in establishment i. The regressions include the same set of control variables $\mathbf{x}$ used in Tables 5 and 6. Table 10 reports the results of these regressions for a variety of potential human resource issues.

There are significant differences in the raw proportion reporting personnel problems between establishments which used STW and WTA in 2009 and those which did not. Some of these differences are quite large. For example, 36 % of establishments which used WTA in 2009 reported that there would be a "high burden from wage costs" in future years, compared to 25 % of establishments which did not

Table 10 Probability of human resource problems in future years

	Did not use STW in 2009	Used STW in 2009	Raw difference	OLS difference
(a) Staff level too high	0.049	0.108	0.060***	0.025**
(b) High staff turnover	0.041	0.061	0.020***	0.009
(c) Difficulties in hiring qualified workers	0.351	0.523	0.172***	0.015
(d) Staff shortages	0.079	0.115	0.036***	0.027**
(e) High burden from wage costs	0.250	0.360	0.110***	0.039**
(f) Other personnel problems	0.033	0.049	0.016***	0.001
	Did not use WTA in 2009	Used WTA in 2009	Raw difference	OLS difference
(a) Staff level too high	0.041	0.073	0.032***	0.005
(b) High staff turnover	0.030	0.057	0.027***	0.004
(c) Difficulties in hiring qualified workers	0.270	0.486	0.216***	0.050***
(d) Staff shortages	0.061	0.108	0.046***	0.005
(e) High burden from wage costs	0.220	0.313	0.093***	0.039***
(f) Other personnel problems	0.030	0.041	0.011***	0.010**

Dependent variable is the answer to the question "What kind of problems with human resources management do you expect for your establishment/office during the next 2 years?". OLS difference are estimates of β and γ from Eq. (6), and include controls for establishment size, sector, location as well as the explanatory variables used in Tables 5 and 6. ***, ** and * indicate significance at the 1%, 5% and 10% level respectively.

use WTA in 2009. In fact in every case the proportion of establishments reporting problems is greater in the treatment group. In some cases these differences seem incompatible. For example, a higher proportion of STW establishments report that they have too high a staff level and that wage costs are too high, but at the same time report that they face staff shortages and difficulties in hiring qualified workers. This seems to suggest a mismatch between establishments' desired levels of skilled workers and their actual employment.

However, most of these raw differences are accounted for by differences in observable characteristics between the treatment and control plants. After controlling for establishment size, sector and location and measures of performance, STW and WTA establishments are both significantly more likely to report a high burden from wage costs, and STW establishments are significantly more likely to report too high a staff level. These results are only suggestive however, since we have only controlled for observable selection into the treatment, and the information is only available in a single cross-section. To confirm these results we would need to track establishments' hires and separations of skilled and unskilled workers over the post-crisis period.

8 Conclusions

Various recent studies have suggested a host of potential explanations for the resilience of the German labour market during the 2008–2009 crisis. However, there is no clear consensus. While some authors have argued that the use of STW was key, others have pointed out that, since the use of short-time work was not that much greater than in recessions in the 1970s and 1980s, the use of short-time work cannot explain the "missing" fall in employment Burda and Hunt (2011).

In this chapter we used a simple descriptive methodology which relates output shocks and job flows to hires and separations for a large panel survey of German establishments. This methodology sheds light on many of the proposed explanations for the resilience of German establishments to the crisis, in particular the role of various institutional arrangements intended to promote workplace flexibility.

Our main findings are as follows. First, establishments in the panel survey did experience a significant fall in sales during the crisis, but the fall in employment was much smaller. This is entirely consistent with the aggregate picture and gives us confidence that the survey is representative. Establishments which reported being directly affected by crisis (which we call "crisis establishments") were more likely to be in high-tech manufacturing sectors, and were more likely to be part of a larger organisation.

Second, the short-run relationship between sales shocks and employment is relatively weak. Negative shocks to sales translate into only small changes in employment, leading by definition to a reduction in labour productivity. However, the estimated response to sales shocks is quite stable over time: there is no evidence of greater labour hoarding in the most recent downturn. In fact, crisis establishments exhibited a *larger* employment response to a given output shock than non-crisis establishments.

Third, the use of STW schemes is very strongly associated with contemporaneous falls in sales, more so than other policy measures such as WTA and PECs. Establishments using STW during the crisis did experience much larger falls in labour productivity than non-STW establishments, but this was largely due to the larger falls in output, not because of smaller falls in employment. In fact, STW establishments had significantly larger falls in employment. We believe that despite conditioning on the change in sales, this result reflects additional (unobserved) negative shocks experienced by STW establishments. We also find no evidence that WTA played an important role in preventing layoffs. This is likely because the size of WTA surpluses at the advent of the global crisis were too small to have any significant buffer effect.

Finally, the fall in sales experienced by STW establishments in 2009 was probably very short-lived, because there was a strong rebound in managers' expectations of performance in 2010. Indeed, managers in STW establishments had far more positive expectations for sales growth in 2010 than those in non-STW establishments. This itself supports the idea that STW was *not* being used by establishments with longer-term structural weaknesses.

Appendix 1: Additional Table

Appendix 2: Questions Used in the IAB Establishment Panel on Worker Turnover

The following questions are used to determine hires and separations:

1. Did you recruit staff in the first half of <current year>?
2. Please indicate the total number of workers recruited.
3. Did you register any staff leaving your establishment/office in the first half of <current year>?
4. Please indicate the total number of workers who left your establishment.

Respondents are also asked to distribute the total number of employees who left among the following categories:

1. Resignation on the part of the employee
2. Dismissal on the part of the employer
3. Leaving after termination of the in-company training
4. Expiration of a temporary employment contract
5. Termination of a contract by mutual agreement
6. Transfer to another establishment within the organization
7. Retirement after reaching the stipulated pension age
8. Retirement before reaching the stipulated pensionable age
9. Occupational invalidity/ disability
10. Other

Table 11 IAB establishment panel: selected sample

	Number of establishments	West Germany	East Germany	Employment (unweighted)	Employment (weighted)	Employment (weighted)
Unbalanced sample						
1993	2,913	2,844	69	532	15	513
1994	3,006	2,930	76	457	15	487
1995	3,061	2,988	73	418	15	506
1996	5,793	2,941	2,852	257	14	520
1997	6,279	2,899	3,380	214	14	566
1998	6,580	2,946	3,634	199	14	487
1999	6,985	2,956	4,029	175	13	423
2000	10,405	6,096	4,309	138	13	506
2001	11,594	7,057	4,537	133	13	462
2002	11,404	7,201	4,203	128	13	373
2003	11,975	7,349	4,626	114	13	464
2004	11,842	7,324	4,518	126	13	463

(continued)

Table 11 (continued)

	Number of establishments	West Germany	East Germany	Employment (unweighted)	Employment (weighted)	Employment (weighted)
2005	12,004	7,381	4,623	127	13	454
2006	11,736	7,172	4,564	120	13	526
2007	12,087	7,453	4,634	109	14	506
2008	11,987	7,251	4,736	106	14	410
2009	12,099	7,394	4,705	101	14	499
2010	12,296	7,513	4,783	93	14	553
Balanced panel 2000–2010						
2000	2,002	900	1,102	108	19	365
2001	2,002	900	1,102	109	19	368
2002	2,002	900	1,102	107	19	362
2003	2,002	900	1,102	105	18	362
2004	2,002	900	1,102	103	18	355
2005	2,002	900	1,102	102	18	349
2006	2,002	900	1,102	102	18	343
2007	2,002	900	1,102	103	19	342
2008	2,002	901	1,101	104	19	347
2009	2,002	901	1,101	102	18	335
2010	2,002	901	1,101	101	18	329

The small number of establishments in East Germany before 1996 are establishments located in West Berlin. The unbalanced panel is weighted by cross-section weights, the balanced panel is weighted by longitudinal weights

References

Abowd, J., & Kramarz, F. (1999). The analysis of labor market data using matched employer-employee data. In O. Ashenfelter, & D. Card (Eds.), *Hanbook of labor economics* (Vol. 3b, pp. 2629–2701, chapter 40). Amsterdam: Elsevier.

Abraham, K., & Houseman, S. (1994). Does employment protection inhibit labor market flexibility? Lessons from Germany, France and Belgium. In R. Blank (Ed.), *Social protection versus economic flexibility: Is there a trade-off?* (pp. 59–94). Chicago: University of Chicago Press.

Arpaia, A., Curci, N., Meyermans, E., Peschner, J., & Pierini, J. (2010). *Short time working arrangements as response to cyclical fluctuations*. European Commission Economic and Financial Affairs Occasional Papers 64.

Bellmann, L., Gerlach, K., & Meyer, W. (2008). Company-level pacts for employment. *Journal of Economics and Statistics, 228*, 533–553.

Bellmann, L., & Gerner, H.-D. (2011a). Further training and company-level pacts for employment in Germany. *Journal of Economics and Statistics, 232*, 98–115.

Bellmann, L., & Gerner, H.-D. (2011b). Reversed roles? Wage and employment effects of the current crisis. *Research in Labor Economics, 32*, 181–206.

Bellmann, L., & Gerner, H.-D. (2012). Company-level pacts for employment in the global crisis 2008/2009: First evidence from representative German establishment-level panel data. *International Journal of Human Resource Management, 23*, 3375–3396.

Bellmann, L., Gerner, H.-D., & Upward, R. (2011). *Job and worker turnover in German establishments*. IZA Discussion Paper 6081.

Boeri, T., & Bruecker, H. (2011). Short-time work benefits revisited: Some lessons from the Great Recession. *Economic Policy, 68*, 697–765.

Boeri, T., & van Ours, J. (2013). *The economics of imperfect labor markets* (2nd ed.). Princeton: Princeton University Press.

Bogedan, C., Brehmer, W., & Herzog-Stein, A. (2009). *Betriebliche Beschäftigungssicherung in der Krise*. Kurzauswertung der WSI-Betriebsrtebefragung 2009.

Bohachova, O., Boockmann, B., & Buch, C. (2011). *Labour demand during the crisis: What happened in Germany*? IZA discussion paper 6074.

Boysen-Hogrefe, J., & Groll, D. (2010). The German labour market miracle. *National Institute Economic Review, 214*(1), R38–R50.

Brenke, K., Rinne, U., & Zimmerman, K. (2011). *Short-time work: The German answer to the great recession*. IZA discussion paper 5780.

Burda, M., & Hunt, J. (2011). *What explains the German labour market miracle in the great recession*? Paper presented at the Brookings Papers conference.

Cahuc, P., & Carcillo, S. (2011). *Is short-time work a good method to keep unemployment down*? CEPR Discussion Paper No. 8214.

Cameron, A. C., & Trivedi, P. (2009). *Microeconometrics using Stata*. Stata Press.

Crimmann, A., Wießner, F., & Bellmann, L. (2010). *The German work-sharing scheme: an instrument for the crisis*. International Labour Organisation Conditions of work and employment working paper series no. 25.

Davis, S., Faberman, R. J., & Haltiwanger, J. (2011). *Labor market flows in the cross-section and over time*. NBER working paper 17294.

Deeke, A. (2005). *Kurzarbeit als Instrument betrieblicher Flexibilität, Ergebnisse aus dem IAB Betriebspanel*. IAB Forschungsbericht 12/2005.

Ellguth, P., & Kohaut, S. (2008). Ein bund fürs Überleben? Einflussfaktoren für die vereinbarung betrieblicher Bündnisse zur Beschäftigungs- und Standortsicherung. *Industrielle Beziehungen, 15*(3), 1–24.

Fischer, G., Janik, F., Müller, D., & Schmucker, A. (2009). The IAB Establishment Panel—things users should know. *Schmollers Jahrbuch, 129*(1), 133–148.

Hijzen, A., & Venn, D. (2011). *The role of short-time work schemes during the 2008–09 recession*. OECD social, employment and migration working papers No. 115.

Hübler, O. (2005). Sind betriebliche Bündnisse für Arbeit erfolgreich? *Journal of Economics and Statistics (Jahrbuecher für Nationaloekonomie und Statistik), 225*(6), 630–652.

Klinger, S., Rebien, M., Heckmann, M., & Szameitat, J. (2011). Did recruitment problems account for the German job miracle? *International Review of Business Research Papers, 7*(1), 265–281.

Möller, J. (2010). The German labor market response in the world recession—de-mystifying a miracle. *Journal for Labour Market Research (Zeitschrift für Arbeitsmarkt Forschung), 42*(4), 325–336.

Morley, J., Sanoussi, F., Biletta, I., & Wolf, F. (2009). *Comparative analysis of working time in the European Union*. Mimeo, European Foundation for the Improvement of Living and Working Conditions.

OECD (2012). What makes labour markets resilient during recessions? In *OECD Employment Outlook 2012* (pp. 53–107, chapter 2). Paris: OECD.

Seifert, H. (2005). Flexibility through working time accounts: reconciling economic efficiency and individual time requirements. In A. Hegewisch (Ed.), *Working time for working families: Europe and the United States*, Friedrich-Ebert-Stiftung.

Sisson, K., & Martin Artiles, A. (2000). *Handling restructuring. a study of collective agreements dealing with employment and competitiveness*. Luxembourg: Office for Official Publications of the European Communities.

Van Audenrode, M. (1994). Short-time compensation, job security and employment contracts: Evidence from selected OECD countries. *Journal of Political Economy, 102*(1), 76–102.

Venn, D. (2009). *Legislation, collective bargaining and enforcement: updating the OECD employment protection indicators*. OECD Social, Employment and Migration Working Papers No. 89.

Complex Network Analysis in Socioeconomic Models

Luis M. Varela, Giulia Rotundo, Marcel Ausloos, and Jesús Carrete

Abstract This chapter aims at reviewing complex network models and methods that were either developed for or applied to socioeconomic issues, and pertinent to the theme of New Economic Geography. After an introduction to the foundations of the field of complex networks, the present summary adds insights on the statistical mechanical approach, and on the most relevant computational aspects for the treatment of these systems. As the most frequently used model for interacting agent-based systems, a brief description of the statistical mechanics of the classical Ising model on regular lattices, together with recent extensions of the same model on small-world Watts–Strogatz and scale-free Albert-Barabási complex networks is included. Other sections of the chapter are devoted to applications of complex networks to economics, finance, spreading of innovations, and regional trade and developments. The chapter also reviews results involving applications of complex networks to other relevant socioeconomic issues, including results for opinion and citation networks. Finally, some avenues for future research are introduced before summarizing the main conclusions of the chapter.

L.M. Varela (✉)
Grupo de Nanomateriais e Materia Branda, Departamento de Física da Materia Condensada, Universidad de Santiago de Compostela, Campus Vida s/n, 15782 Santiago de Compostela, Spain
e-mail: luismiguel.varela@usc.es

G. Rotundo
Department of Methods and Models for Economics, Territory and Finance, Sapienza University of Rome, via del Castro Laurenziano 9, 00161 Rome, Italy
e-mail: giulia.rotundo@gmail.com

M. Ausloos
Rés. Beauvallon, rue de la Belle Jardiniére, 483/0021, 4031 Liège Angleur, Euroland, Belgium

eHumanities Group, Royal Netherlands Academy of Arts and Sciences, Joan Muyskenweg 25, 1096 CJ Amsterdam, The Netherlands (previously at GRAPES, ULG, Liege, Belgium)
e-mail: marcel.ausloos@ulg.ac.be

J. Carrete
Grupo de Nanomateriais e Materia Branda, Departamento de Física da Materia Condensada, Universidad de Santiago de Compostela, Campus Vida s/n, 15782 Santiago de Compostela, Spain

LITEN, CEA-Grenoble, 17 rue des Martyrs, BP166, 38054, Grenoble, Cedex 9, France
e-mail: jcarrete@gmail.com

P. Commendatore et al. (eds.), *Complexity and Geographical Economics*,
Dynamic Modeling and Econometrics in Economics and Finance 19,
DOI 10.1007/978-3-319-12805-4_9

1 Introduction

The foundation of the field of network topology dates back to the eighteenth century with the seminal work in graph theory of Euler devoted to the celebrated problem of Königsberg bridges (Euler 1736), and includes several important contributions in the last two centuries like Cayley trees (Cayley 1889) and the theory of random graphs (Erdös and Rényi 1959). However, it was not until the late 1990s that complex networks with specific structural features valid for the description of short path lengths, highly clustered (Watts and Strogatz 1998), and even heterogeneous networks (Barabási and Albert 1999) were introduced, which opened what could be called the contemporary era of network theory. The amount of papers published since then has never ceased to increase exponentially as network theory started to be applied to fields like physics, biology, computer science, sociology, epidemiology, and economics among others. It is now recognized that a network is always the skeleton of any complex system, so it is by no means an exaggeration to say that network theory has become one of the cornerstones of the theory of complex systems. All this activity has been excellently reviewed in several occasions during the last decade (see for e.g. Albert and Barabási 2002; Newman 2003; Pastor-Satorras et al. 2003; Boccaletti et al. 2006; Newman et al. 2006), and it is now widely accepted that the dynamics of many complex systems corresponds to emergent phenomena associated to the large scale fluctuations of some real network.

The effects of the topological properties of networks on dynamical processes is also a matter of intense research inside the field of complex networks. Some of these dynamical processes are the evolution of the network itself, spreading processes in agent-based systems (epidemics in a population, rumor spreading), opinion formation, cultural assimilation, voting processes, or decision making on competing for limited resources (for a review, see Boccaletti et al. 2006). Specifically, the description of the evolution of the network is very important on itself, as real networks evolve in time. For that, one has to follow the evolution of networks, through the number of vertices, the number, weight and direction of links, and through other characteristic quantities mentioned below, as seen in Boccaletti et al. (2006), Lambiotte and Ausloos (2006b, 2007b) and Ausloos and Lambiotte (2007b).

In the field of economics and economic geography, a complete understanding of economic dynamics requires the understanding of its agent-based underlying structure and the interactions that give rise to the observed emergent spatial and temporal organizations, which are definitively more than the sum of its individual components (for further details see e.g. Ausloos et al. 2014). The view of the economy as an evolving complex agent-based system, hereafter identified with a network, is currently gaining consensus among the scientific community (see e.g. Namatame et al. 2006; Barabási 2003; Dorogovtsev and Mendes 2003). On the basis of this network-based structural view, the application of the methods of physics of complex systems (statistical physics, nonlinear physics, and so forth) has allowed to gain new insight on the economic realm, even leading to the foundation of a new branch of Physics itself, the so called Econophysics.

The purpose of this chapter is to summarize some applications that complex network theory has found in economic and social studies. Before reviewing some applications of complex networks to economic issues (financial markets, spreading of innovations, economic geography, regional trade and development, ...), the main results that have been developed up to now in the field of statistical mechanics of complex networks and their computational analysis will be briefly summarized. The chapter is closed with complex-network-based descriptions of other social networks, a brief description of future trends in the field of socio-economic applications of complex networks and our conclusions.

2 Summary of Statistical Mechanics of Complex Networks

As mentioned previously, in an attempt to reproduce the success in describing regular systems (solid state physics, phase transitions and so forth), statistical mechanics has been applied to heterogeneous systems through the formalism of complex networks, representing the constituents by means of vertices and their interactions by a set of edges. The description of these objects involves their topology and dynamic evolution, as well as different dynamic processes that take place over them. One consideration consists in tying the structure of the network and its intrinsic dynamics (Newman 2003). Another concerns the structural changes in the network due to the dynamical processes themselves. As we have previously mentioned, several excellent reviews have been published during the past decade on the structure and dynamics of complex networks, as well as on dynamic processes that take place in these topological objects. However, for the interest of the reader, in this section the main topics of the field will be briefly summarized.

From a formal point of view a network (graph) is a pair (V, E), where V is a set of nodes (vertices), and E is a set of links (edges), which are identified by two nodes that represent the source and the end of the link (edge). The most used method for the representation of networks with N nodes is the *adjacency matrix* $A = (a_{ij}) \in \mathbb{R}^{N\times N}$, whose rows and columns are the nodes of the network, and whose term $a_{ij} > 0$ corresponds to the weight of the link from node i to node j. The absence of links is given by zero elements $a_{ij} = 0$. Therefore, properties of the network are reflected in the properties of the adjacency matrix. Network links are called directed if matrix A is not symmetric, which means that there exists at least one pair of indices i, j such that $a_{ij} \neq a_{ji}$. This includes as particular case graphs with upper triangular adjacency matrices, i.e. such that $a_{ij} > 0$ and $a_{ji} = 0$. In graph theory, such networks are named *directed acyclic graphs*. Undirected networks are represented by symmetric matrices ($A = A^T$). A source i is a node having only outgoing links (i.e. $\forall k \in V$ there are no pairs $(k, i) \in E$). By contrast, a sink j is a node having only incoming links (i.e. $\forall k \in V$ there are no pairs $(j, k) \in E$). A path from node i_{h_1} to node i_{h_k} in the network is a sequence of nodes $i_{h_1}, i_{h_2}, \cdots, i_{h_k}$

such that $a_{i_{h_j} i_{h_j+1}} \neq 0$, and can be detected through the power of the adjacency matrix A^{h_k}.

Paths are relevant for studying diffusion processes as well as the relevance of nodes. A network is called *(strongly) connected* if a path exists between any pair of nodes, considering the direction of the links, and it is termed *(weakly) connected* if a path exists among any pair of nodes when ignoring the direction of the links, which requires that the adjacency matrix be symmetric. In some applications the weight assigned to each link is 1, since the presence of the edge is relevant, but not its specific weight. In this case the matrix A is symmetrized setting $a_{ji} = 1$ for all nodes i, j such that $a_{ij} = 1$. In applications of networks to problems in which the weight of the link is relevant, whether symmetrization is applied depends on the objectives.

The main difference between network theory and graph theory lies in their targets. Naturally, algorithms first developed in the field of graph theory are useful and currently used also for complex networks and in other fields related to dynamical systems and the discretization of maps, like symbolic images (Avrutin et al. 2006; Rotundo 2013).

2.1 Main Measures for Complex Networks

Networks can be classified according to a relatively large number of criteria. Taking into account the distribution of the number of links per node, networks can be classified as purely random networks (Poisson distributed), exponentially distributed small-world networks, and scale-free networks. According to the directionality of contacts they can be classified as directed or undirected graphs. According to the heterogeneity in the capacity and the intensity of the connections, networks can be classified into weighted or unweighted networks depending on whether different weights are associated to their edges. Sparse and fully connected networks differ in the fraction of interconnected nodes. Finally, depending on their time evolution, networks can be classified into static and evolving.

The description of the topology of complex networks is based on several concepts and parameters that measure different features of these topological structures and macroscopic characteristics of the networks (a more detailed treatment can be found in da Costa et al. 2007). The most important of these are:

1. Average path length: no proper metric space can be defined for complex networks (usually hidden metric spaces are defined, see e.g. Serrano et al. 2008), and the (chemical) distance between any two vertices l_{ij} is defined to be the number of steps from one point to the other following the shortest path. In many real networks, the average distance between two nodes, i.e. the average path-length, $< l >$, is relatively small as compared to the total number of nodes in the network. In fact, in a regular lattice, a topological structure which can be generated from a basis for the vector space by forming all linear combinations

with integer coefficient and where all the nodes are connected to the same number of neighbors, the average path length scales with the number of nodes in the network, N, as $< l > \sim \sqrt{N}$, while in a small-world network with long-distance shortcuts, $< l > \sim \log N$, so the separation between any two nodes is usually very small. This property is behind the small-world concept first studied by Milgram in his 1967 seminal paper (Milgram 1967), and it is by no means included in conventional regular lattices. The connectivity can also be measured by means of the diameter of the graph, d, defined as the maximum distance between any pair of its nodes. Djikstra algorithm (Djikstra 1959) is the most often used for this calculation.

2. Degree distribution, $p(k)$, measuring the probability that a given node has k connections to other nodes. This is probably the most important property of networks, and it is behind their classification as exponentially distributed Watts–Strogatz (WS) networks ($p(k) \sim e^{-\alpha k}$) and scale-free Barabási–Albert (BA) networks ($p(k) \sim k^{-\gamma}$). For purely random Erdös–Renyi networks, $p(k)$ is a binomial distribution (so also belonging to the exponentially distributed class),

$$p(k) = \binom{N}{k} p^k (1-p)^{N-k} \tag{1}$$

Sparse networks are those for which the average degree remains finite when $N \to \infty$, and for real networks, $< k > \ll N$.

3. Clustering: The clustering coefficient c_i of a vertex i is given by the ratio between the number e_i of triads—connected subsets of three network nodes—sharing that vertex, and the maximum number of triangles that the vertex could have. Alternatively, this coefficient is the ratio between the number E_i of edges that actually exist between the k_i neighbors of vertex i and the maximum number $k_i(k_i - 1)/2$. The clustering coefficient provides a measure of the local connectivity structure of the network. This coefficient usually takes large values in social networks, contrary to what happens in random graphs (Albert and Barabási 2002). The average cluster coefficient is given by $< c >= \sum_i c_i / N$ and the clustering spectrum by $< c(k) >= \sum_i \delta_{kk_i} c_i / N p(k)$, where N is the number of vertices in the network, and $p(k)$ is the degree distribution defined below. A graph is considered to be small-world if its clustering coefficient is considerably greater than that of a random graph built on the same node set and the average path length is approximately the same as that of the corresponding random graph. One may point out that the clustering coefficient of triangular Erdös–Renyi networks, i.e. uncorrelated random graphs, is very high by construction, but, contrary to intuition, it is different from 1 in general.
4. The overlapping index (Gligor and Ausloos 2008b) measures the common number of neighbors of the i and j nodes, i.e. how many triads have a common basis. Moreover, a network transitivity is the probability that two neighbors of a node have themselves a link between them. In topological terms, it is a measure of the density of triads in a network.

2.2 Additional Parameters and Concepts

In addition to the ones mentioned above, there are a lot of other parameters and concepts that measure important properties of the topology of complex networks. A thorough treatment of these can be found in more detailed reviews like the ones cited in this chapter or in monographs like for e.g. that in Pastor-Satorras et al. (2003). For brevity, we shall mention just a few of them of particular interest.

1. Centrality of a node (Friedman 1977). Many measures are gathered under the label "Centrality". The node degree, i.e. its number of contacts to other nodes of the network, is one of them. Another one is the so called betweenness of a vertex b_i or an edge b_{ij}, which is the number of shortest paths that pass through the vertex i (edge (i, j)), for all the possible pairs of vertices in the network.
2. Correlations in networks: The correlations usually found in real networks (i.e. the fact that the degrees of the nodes at the ends of a given vertex are not in general independent) are measured by means of the distribution $p(k \mid k') = k' p(k')/ < k >$, representing the conditional probability that an edge that has one node with degree k has a node with degree k' at the other end. In a correlated network, this distribution depends both in k and k', while in an uncorrelated network, it depends only on k'. An alternative measure of correlations is given by the average degree of the nearest neighbors of the vertices of degree k (Barrat et al. 2004),

$$< k_{nn} >= \sum_{k'} k' p(k \mid k') \tag{2}$$

   The network is said to be correlated if this parameter depends on k. Other measures are the closeness centrality or the flow-betweenness centrality. When $< k_{nn} >$ increases with k the network is called assortative, while if $< k_{nn} >$ is a decreasing function of k, the network is called disassortative. Then, the assortativity coefficient (Newman 2002a) measures a network property through the Pearson correlation coefficient of the degrees at either ends of an edge. Finally, one must recall that $p(k \mid k')$ and $p(k)$ are not independent, but, due to the conservation of edges, they are related by a degree detailed balance condition (Boguñá and Pastor-Satorras 2002)

$$kp(k)p(k' \mid k) = k' p(k')p(k \mid k') \tag{3}$$

3. k-shells. A k-shell of a graph G is a connected subgraph of G in which all vertices have degree at least k. Equivalently, it is one of the connected components of the subgraph of G formed by repeatedly deleting all vertices of degree less than k. The k-core of a graph G is the k-shell with the maximum k. The concept of k-shells and k-cores was introduced to study the clustering structure of social networks and to describe the evolution of random graphs; it has also been applied in bioinformatics and network visualization. In Economics

and Finance it has been applied to the corporate ownership network, to the international trade network, and to the network of shareholders (Garas et al. 2010, 2012; Rotundo and D'Arcangelis 2014).

4. Nestedness index (Araujo et al. 2010). It indicates the likelihood that a node is linked to the neighbors of the nodes with larger degrees. The mean topological overlap between nodes (Almeida-Neto et al. 2008) has been introduced to quantify nestedness.

2.3 *Ising Model on Complex Networks*

Several classical problems of statistical mechanics have been now studied using complex networks. In particular, the mean-field solution for the average path length and for the distribution of path lengths in small-world networks have been reported (Newman et al. 2000). On the other hand, the mean-field solution of the Ising model (Ising 1925) on a small-world complex network has been contributed by several authors (Gitterman 2000; Barrat and Weigt 2000; Viana-Lopes et al. 2004; Herrero 2002), and by Bianconi on a BA network (Bianconi 2002) in the first half of the last decade. Viana-Lopes solved the 1D Ising model on a small-world network (Viana-Lopes et al. 2004), and Herrero considered the ferromagnetic transition for the Ising model in small-world networks generated by rewiring 2D and 3D lattices (Herrero 2002). Due to its interest for socioeconomic researchers, some of the main results reported in these contributions are recalled below.

The Ising model originally introduced for the study of ferromagnetism, is the simplest paradigm of order-disorder transitions (Ising 1925). Undoubtedly it represents one of the major milestones in the development of statistical mechanics of interacting systems and phase transitions, and it is at the basis of a plethora of applications and generalizations reported throughout the twentieth century. This is a discrete model in which spins (agents) with two possible states (± 1) are placed in the nodes of a graph (normally a lattice) and are allowed to interact with their nearest-neighbours, being probably the simplest model to exhibit a phase transition. The one dimensional problem was solved by Ising himself in 1925 (Ising 1925), and Onsager reported the exact solution to the 2D problem (Onsager 1944). The model Hamiltonian for the N spins $s_i = \pm 1$ on the nodes of the graph is given by

$$H = -\sum_{i,j} J_{ij} s_i s_j - \sum_i h_i s_i \tag{4}$$

h_i being the local external (magnetic) field, and J_{ij} being nonzero only for those pairs of spins connected by a link. If $J_{ij} > 0$ then parallel orientations of spins are energetically favored (ferromagnetism), while for $J_{ij} < 0$ antiparallel orientations are preferred (antiferromagnetic case). In social physics applications these cases

correspond, respectively, to consensus/dissensus-oriented models. For the nearest neighbours 1D problem, Eq. (4) can be rewritten as

$$H = -J \sum_{i=1}^{N} s_i s_{i+1} - \sum_{i=1}^{N} h_i s_i, \tag{5}$$

with $s_{N+1} = s_1$ as usual for periodic boundary conditions. The canonical partition function—the normalization factor of the probability density in the phase space of the system's microstates—is straightforwardly calculated from the above equation as:

$$\begin{aligned} Z(\beta) &= e^{-\beta H(\{s_i\})} \\ &= \sum_{s_1 \ldots s_N} \prod_{i=1}^{N} e^{\beta h s_i} e^{\beta s_i s_{i+1}} \\ &= \sum_{s_1 \ldots s_N} \prod_{i=1}^{N} \Delta_{s_i s_{i+1}} \end{aligned} \tag{6}$$

where $\beta^{-1} = k_B T$ represents the thermal energy, $\{s_i\}$ a configuration of the spins and $\Delta_{s_i s_{i+1}} = \exp(\beta h s_i/2) \exp(\beta s_i s_{i+1}) \exp(\beta h s_{i+1}/2)$ is the transfer matrix. Thus, using conventional matrix algebra, we can obtain the partition function in terms of the eigenvalues of the matrix Δ, $\lambda_i (i = 1, 2)$, as:

$$\begin{aligned} Z(\beta) &= \mathrm{Tr}\left(\Delta^N\right) \\ &= \lambda_1^N \left[1 + \left(\frac{\lambda_2}{\lambda_1}\right)^N\right] \end{aligned} \tag{7}$$

and the associated Helmholtz free energy per spin (a thermodynamic potential comprising all the relevant thermodynamic information about the system and governing equilibrium and stability at constant temperature and volume), $f(\beta)$, as:

$$\begin{aligned} -\beta f(\beta) &= \lim_{N \to \infty} \frac{\ln Z(\beta)}{N} \\ &= \ln\left(e^{\beta J} \cosh(\beta h) + \sqrt{e^{2\beta J} \sinh^2(\beta h) + e^{-2\beta J}}\right) \end{aligned} \tag{8}$$

Moreover, one can prove (see for example LeBellac 1992 for an elegant treatment of the topic) that spin-spin correlations are given in this model by:

$$\left\langle s_i s_j \right\rangle = e^{-r_{ij}/\xi} \tag{9}$$

where $\xi = a/|\ln\tanh(\beta J)|$, with a being the lattice spacing between spins. It is well-known (LeBellac 1992) that a 1D system exhibits no phase transition in the absence of long-range interactions. In 2D systems this ferromagnetic transition ($J > 0$) is registered, and it was Onsager (1944), who obtained the partition function for the vanishing external magnetic field case, and Yang who calculated the magnetization in the ferromagnetic phase (Yang 1952)

$$M = \left\{1 - \left[\sinh\left(\log(1+\sqrt{2})\frac{T_c}{T}\right)\right]^{-4}\right\}^{1/8} \tag{10}$$

Here $k_B T_c = 2J/\log(1+\sqrt{2}) \simeq 2.27J$ is the critical temperature where the ferromagnetic (consensus) transition takes place.

Contrarily to their physical homologues, where interactions are usually of limited range, in social systems long-range connections between agents are very frequently registered. In order to preserve an Ising-based descriptions of these systems, it is necessary to generalize the formalism so as to include the existence of long-range correlations between agents, and that is most conveniently done by means of complex networks.

The solutions of the 1D Ising model in complex networks has been reported by Viana-Lopes et al. on a small-world network (Viana-Lopes et al. 2004), and by Bianconi on a scale-free network (Bianconi 2002), the latter in the mean-field approximation. In the former work, the authors exactly solved a one-dimensional Ising chain with nearest neighbor interactions and random long-range interactions, obtaining a phase transition of a mean-field type. The authors considered a 1D lattice in which the bonds are rewired at random with a probability p in a Watts–Strogatz fashion, and they used a Hamiltonian where they allowed the existence of different short-range (chain, J) and long-range (I) interactions,

$$H = -J\sum_{i=0}^{N} s_i s_{i+1} - I\sum_{ij\in S} s_i s_j - h\sum_{i=1}^{N-1} s_i \tag{11}$$

where the set S includes the $N_b = Np$ shortcut pairs of nodes connected by a long-range connection. Even for small values of the shortcut probability p, a dramatic increase in the connectivity of the network is registered, which has a deep influence in the thermodynamics of the Ising problem. Particularly, Viana-Lopes et al. (2004) proved that a phase transition exist in this 1D problem with a transition temperature given by

$$t_J^d(1+2t_I) = 0 \tag{12}$$

where $d = 1/2p$, $t_J = \tanh(\beta J)$ and $t_I = \tanh(\beta I)$. The authors analyzed the behavior of this temperature in the limits of shortcut bonds stronger than chain bonds

($pI \to \infty$ for any finite p) and of chain bonds much stronger than shortcut bonds ($pI \to 0$) and obtained:

$$T_c = \frac{2J}{\ln(1/p \ln 3)} \qquad pI \to \infty$$
$$T_c = \frac{2J}{\ln\{J/[pI \ln(J/pI)]\}} \qquad pI \to 0 \tag{13}$$

On the other hand, the Ising model on a BA network has been solved in the mean-field approximation (Bianconi 2002), as previously mentioned. The author showed that the mean-field solution of the Ising model in this type of network can be treated as a Mattis model, a simple solvable model of the spin glass in which Ising spins interact via unfrustrated random exchange interactions (Mattis 1976). The author considered a BA network of N spins constructed iteratively with the constant addition of new nodes with m connections and a Hamiltonian

$$H = -J \sum_{i,j} \epsilon_{ij} s_i s_j - \sum_{i=1}^{N} h_i s_i, \tag{14}$$

Here $\epsilon_{ij} = \langle A_{ij} \rangle = k_i k_j / 2mN$ is the average of the adjacency matrix over many copies of the network (Bianconi 2002). The mean-field solution of the Hamiltonian for the order parameter S is

$$S = \frac{1}{2mN} \sum_{i=1}^{N} k_i \langle s_i \rangle$$
$$= \frac{1}{2mN} \sum_{i=1}^{N} k_i \tanh[\beta (Jk_i S + h_i)] \tag{15}$$

where one can see that the effective mean-field acting on a spin is determined not only by the external field and the interaction strength, as in conventional Ising model, but also by the connectivity of the network nodes. The above equation resembles that of the Mattis model with the substitution of the quenched random variables ξ_i in Mattis (1976) by the node degree k_i. Bianconi was able to prove from the above that the effective critical temperature is given by

$$T_c = \frac{mJ}{2} \ln(N), \tag{16}$$

i.e. it increases linearly with the interaction J and logarithmically with the number of nodes in the system, in agreement with previously reported numerical simulations (see Bianconi 2002 for further details).

Finally, it is worth mentioning that the existence of phase transitions in 1-D systems with long-range interactions has been proved by means of Monte Carlo simulations by Pekalski, who demonstrated that even a small fraction of long-distance shortcuts induces ordering of the system at finite temperatures and provided the dependence of the magnetization and the critical temperature on the concentration of the small world links (Pekalski 2001).

2.4 A Special Case: Bipartite Networks Reductions

For our present purposes, it must be here emphasized that when discussing interacting economic entities, it is paramount to discriminate N-body correlations that are intrinsic N-body interactions from those that merely develop from lower-order interactions, like the 2-body interactions of the Ising model. This issue is directly related to a well-known problem in complex network theory, i.e. the projection of bipartite networks composed of two kinds of nodes, onto unipartite networks, i.e., composed of one kind of node. This property of bipartiteness is a special case of disassortativity. A network is called bipartite if its vertices can be separated into two sets such that edges exist only between vertices of different sets (Lambiotte and Ausloos 2005a). Bipartite networks are well known in graph theory and operations research, where the delivery problem from N sources to M sinks is well studied, and finds a first application in Economics to the problem of finding the optimal supply of goods (in the N sources) to accomplish the demand function of the M consumers (the M sinks) (Hotelling 1929). The model may vary to consider the optimal location and geographical distance among economic activities (the N sources) and the customers (the M sinks). Further recent applications to financial networks relate the set of N companies to the set of their M directors, who are persons, so the two sets are naturally describing two very different categories. This is naturally a bipartite graph, and there is a link among a company and a person if he/she is in the administrative board of the company. This network can be represented through a matrix $A \in R^{N \times M}$ that is the starting point for the study of ties among companies given by the presence of the same directors in their boards (AA^T), or the connections among persons due to belonging to the same boards ($A^T A$) (Caldarelli et al. 2004; Bertoni and Randone 2006; Rotundo and D'Arcangelis 2010a; Grassi 2010; Croci and Grassi 2013).

This formalism and a coarse graining description of bipartite networks from a statistical mechanics approach can be found in Lambiotte and Ausloos (2005a,b, 2006c) and Newman (2002b). Formally, the bipartite structure of e.g. quantities or prices *vs*. producers may be mapped exactly on the *vector* of matrices $\mathscr{M}$ defined by:

$$\mathscr{M} = [M^1_{a_1}, M^2_{a_1 a_2}, M^3_{a_1 a_2 a_3}, \ldots, M^{n_P}_{a_1 \ldots a_{n_P}}] \tag{17}$$

where $\mathbf{M}^j$ is a square n_P^j matrix that accounts for all quantities (at some price) j produced by producer P. For example, $M^1_{a_1}$ and $M^2_{a_1 a_2}$ represent respectively the total number of goods produced by a_1 alone, and the total number of goods produced by the pair (a_1, a_2).

It is important to point out that the vector of matrices $\mathscr{M}$ describes the bipartite network without approximation, and that it reminds of the Liouville distribution in phase space of a Hamiltonian system. Accordingly, a relevant macroscopic description of the system relies on a coarse-grained reduction of its internal variables. The simplest reduced matrix is the single producer matrix:

$$R^1_{a_1} = M^1_{a_1} + \sum_{a_2} M^2_{a_1 a_2} + \sum_{a_2} \sum_{a_3 < a_2} M^3_{a_1 a_2 a_3} + \ldots + \sum_{a_2} \ldots \sum_{a_j < a_{j-1}} M^j_{a_1 \ldots a_j} + \ldots \tag{18}$$

that is a vector whose elements $R^1_{a_j}$ denote the total number of goods produced by a_j. The second order matrix:

$$R^2_{a_1 a_2} = M^2_{a_1 a_2} + \sum_{a_3} M^3_{a_1 \ldots a_3} + \ldots + \sum_{a_3} \ldots \sum_{a_j < a_{j-1}} M^j_{a_1 \ldots a_j} + \ldots \tag{19}$$

Its elements represent the total number of quantities produced by the pair (a_1, a_2).

Remarkably, this matrix reproduces the usual projection method and obviously simplifies the bipartite structure by hiding the effect of higher order interactions. One may next discriminate between different types of triangles and discuss, e.g. the interplay between producers at the node degree level. Economic directed and weighted networks such as payment networks (see Bougheas and Kirman 2015) needs to be further explored.

3 Computational Description of Complex Networks

The representation of graphs and networks as data structures in computer memory is by now well established, with several comprehensive and efficient implementations openly available (Siek et al. 2001). The representation determines both the storage requirements and the efficiency of common operations on networks, and must be chosen accordingly. It must be noted that the available formats differ mostly in how edges are stored, as the most appropriate structure for storing data about vertices is relatively independent of the topology.

A particularly simple data structure for storing graphs is the adjacency matrix $A = (a_{ij})$ with matrix elements equal to the number of edges in the graph. Among adjacency matrices, the less complex are those of simple undirected graphs with no edge weights, where the diagonal elements are always zero, $a_{ij} = a_{ji}$ and each

element is 1 if the corresponding edge exists and 0 otherwise. However, when edge directionality is introduced the symmetry constraint must be abandoned, and if edges have weights they can be used as matrix entries. The main virtue of adjacency matrices, beyond the straightforwardness of their implementation, is the efficiency of adding/removing edges and checking for their existence. On the other hand, they are particularly poorly suited for representing social networks, which tend to be sparse: in a naive implementation of an adjacency matrix for a graph with N vertices, most of the N^2 entries will be zero. However, these matrices can still be used as long as they are stored in one of the formats commonly employed for sparse arrays, such as a coordinate list (list of tuples (i, j, a_{ij}) containing only those for which $a_{ij} \neq 0$) or a dictionary using the (i, j) coordinates of non-zero elements as keys (Pissanetzky 1984). Alternative, specific formats exist for graphs. For instance, in an adjacency list the set of neighbors is stored along with the rest of the data for each vertex, which keeps vertex addition and removal very efficient while ensuring that storage space scales only linearly with N.

It is not often that one finds the opportunity, or even the need, to model a real-world social network in a completely detailed fashion. Although such studies might be useful to assist political decisions, most theoretical approaches are commonly focused on exploring general phenomena, for which such overfitted models would be of little help. Instead, ensembles of random networks are built that capture just the essential features of the real networks. Of interest for social models is the fact that in most generating algorithms the structure of a network is a phenomenon that emerges from its growth dynamics. For instance, maximally random networks, where each of the $N(N-1)/2$ edges has the same probability (p) of being present regardless of the degrees of the vertices it joins, can be created using the classical Gilbert algorithm (Gilbert 1959), in which edges are added at random to an initially disconnected graph. The results are equivalent to those of the Erdös–Renyi model (Erdös and Rényi 1959): networks that have no loop with lengths much shorter than the size of the graph. These local tree-like loops give rise to the small-world property of purely random graphs, in which the average distance between nodes grows only proportionally to the logarithm of N. Another very important phenomenon observed in these simple models is the emergence of a giant connected component, comprising a finite fraction of all nodes even in the infinite-network limit, when $p > 1$.

Both of the aforementioned properties have been observed repeatedly in real-world social networks. One of the most striking examples is a recent comprehensive study of the structure of the Facebook "friend" network, with 721 million users (Ugander et al. 2011), 99.91 % of them in a single connected component and an average distance between nodes of just 4.71. However, the limitations of Erdös–Renyi for describing real networks are also readily encountered when looking at the local environment of each vertex. The distribution of node clustering coefficients (number of links that exist among the neighbours of a degree-k node, normalized to the maximum number $k(k-1)/2$ that could exist) or the network clustering

coefficients (fraction of the total possible 3-loops, or triads, actually present in the graph) tends to zero in the infinite-network limit and underestimates observed values by two to three orders of magnitude. This is a reflection of the fact that in real-life processes such as interactions between individuals, links are more likely to be formed with other nodes in the local environment as opposed to distant ones, a bias not taken into account at all in simple random models. Another trivial limitation is that the Bernoulli processes used to build these families of graphs result in Poisson-like exponential distributions, while the shape of experimental distributions is often found to be less quickly decaying, and even scale-free.

The first problem is tackled by modified construction algorithms such as the celebrated Watts–Strogatz model (Watts and Strogatz 1998), that takes a regular network with a ring topology (high clustering coefficient, long mean distance between nodes) as its starting point and proceeds by replacing its edges with random shortcuts with a uniform probability p. The small-world limit is quickly reached even for small values of p, whereas the clustering coefficient decays more slowly as $(1-p)^3$. Hence, Watts–Strogatz networks interpolate between the desirable properties of regular graphs and the Gilbert model. However, their degree distribution is still exponential for large k.

In contrast, the Barabási–Albert model (Albert and Barabási 2002) for network growth dynamics achieves a scale-free degree distribution with a $\propto k^{-3}$ fat tail by using a preferential attachment mechanism: the starting point is a small and fully connected network and at each step a new node is added and connected to m of the existing ones with probabilities proportional to their degrees. The very connected nodes present in the final network result in mean distances between nodes even shorter than those in small-world networks. Other functional choices for representing the preference for attachment with well-connected nodes are possible, and as long as they are asymptotically linear they also result in scale-free degree distributions (Krapivsky et al. 2000), with exponents down to 2. This limit can be overcome by abandoning pure preferential attachment and introducing more drastic measures, such as merging of vertices (Seyed-allaei et al. 2006), into the dynamics of the network. It must be mentioned here that, since real studies are always performed on finite networks, and fat tails may only become clear at high values of k, deciding on whether a real network is scale-free or not can be mathematically and computationally challenging; in fact, different groups working on the same data have occasionally reached opposite conclusions (Cotilla-Sanchez et al. 2012).

Still, none of these models offer the researcher the possibility to computationally build ensembles of networks with a known degree distribution, possibly taken from experiment. Fortunately, it is straightforward to generalize, for instance, the idea behind the Erdös–Renyi model to sample graphs uniformly from the ensemble formed by all those with the desired degree distribution (Newman et al. 2001). The key feature of such models is the fact that, by construction, the degrees of the nodes at either end of the same link are not correlated. More specifically, their joint degree distribution $p(k,k')$ can be factorized as $p_e(k)\, p_e(k')$; here, $p_e(k)$ is the

degree distribution of an end of a randomly chosen edge, related to the node degree distribution $P(k)$ by $p_e(k) = p(k)/k$. When their degree variance is finite, these uncorrelated networks are also locally tree-like, but their clustering coefficients in finite cases approximate those of real networks much better than the ones derived from the Gilbert model. Moreover, it is still possible to derive a general condition for the emergence of a giant connected component in this setting, the Molloy-Reed criterion (Molloy and Reed 1995, 1998), $\langle k^2 \rangle > 2 \langle k \rangle$. This criterion is met by scale-free networks, whose second moment diverges in the infinite limit, which explains their extreme resilience.

Clearly, not even the most general uncorrelated networks can capture all the varieties of real-world structures. Even though the joint distribution $P(k, k')$ is seldom available, the Pearson correlation coefficient between k and k', known in this context as assortativity (r) can be used (Newman 2002a) to discriminate between correlated and uncorrelated networks. Preferential-attachment algorithms such as Barabási–Albert give rise to assortative ($r > 0$) networks, while electrical grids, for instance, are know to be dissortative ($r < 0$) (Cotilla-Sanchez et al. 2012).

A computational study of a network in the time domain will typically start with a network in an initial condition that can be empty, comprehensively sampled from real-world data or generated using one of the algorithms described above and incorporating only the necessary information. It will then proceed through a set of discrete time steps until the desired convergence is achieved. At each step, both the structure of the network and any node or edge variables may change in response to external fields, internal phenomena or interactions through the network. Several points make this kind of simulation very different from models based purely on differential equations, with a longer tradition in physics and economics. First, the computational demand of agent-based models can be formidable when compared with more aggregate treatments of similar systems; thus, automation and parallelization are critical. Even when the network dynamics can be implemented synchronously, with changes only depending on data from previous steps, to avoid race conditions, scaling can be hindered by the need to exchange large amount of data during updates between steps. Moreover, judicious choice of what aggregate variables to track to characterize the evolution of the network (Lambiotte and Ausloos 2006b; Ausloos and Lambiotte 2007b) is of crucial importance to separate signal from noise. In addition to a set of basic descriptors (number of nodes and links, weights and directions if applicable, and so on) and degree and clustering distributions, more problem-specific metrics such as assortativity, betweenness centralities (Friedman 1977), and overlap (Gligor and Ausloos 2008b) and the nestedness indices may be necessary. Finally, given the very specialized nature of network visualization, storage of data should be done in standard formats (HDF5, NetCDF) and the network itself made available in well-documented formats such as GraphML and GML to retain freedom to look at the results from different tools and environments. Third, judicious choice of what aggregate variables to track is of crucial importance to separate signal from noise. Open, scalable, general-purpose visualization tools (Batagelj and Mrva 2003; Bastian et al. 2009) are currently available to ease post-processing work.

4 The Economy as a Complex Adaptive System: Complex-Network-Based Market Models

In the last decades, there has been an increasing amount of effort to conceive the economy as an evolving complex system (see for example Barabási 2003; Tesfatsion 2003; Kirman 1997; Barkley Rosser 1999; Colander et al. 2004; Arthur 2006; Foster 2005; Martin and Sunley 2007; Jackson 2011), meaning that it is a dynamic network of interactions between similar, connected agents which self-organize in order to adapt to a changing environment and maintain the organization of the macrostructure. Many of these approaches employ the complex network formalism—either as a theoretical or a computational tool—for analyzing different perspectives of the economic realm: (a) economics formalism itself, (b) finance, and (c) social networks applied to economics. Below, some of the main contributions in the field are reviewed.

Theoretical and computational agent-based models of economic interactions are being progressively more used by the scientific community to describe the emergent phenomena and collective behavior in economic activities arising from the interaction and coordination of economic agents (Namatame et al. 2006). The complex-network formalism is essential for most of these descriptions. This kind of approach has been used in several problems such as international trade (Bhattacharya et al. 2008), finance (Caldarelli et al. 2013; da Cruz and Lind 2012), globalization (Kali and Reyes 2007; Rodrigue 2013; Xiang et al. 2003; Garlaschelli and Loffredo 2005; Schweitzer et al. 2009) and so forth. Specifically, Schweitzer et al. (2009) analyzed economic networks as tools for understanding contemporary global economy. These authors considered the ability of these objects to model the complexity of the interaction patterns emerging from the incentives and information behind agents' behavior from which metastabilities, system crashes, and emergent structures arise. A detailed characterization of these phenomena requires, according to the authors, "a combination of time-series analysis, complexity theory, and simulation with the analytical tools that have been developed by game theory, and graph and matrix theories".

4.1 Economics: Gross Domestic Product and Other Macroeconomic Indicators

Many contributions have been reported that consider the application of complex networks to the analysis of gross domestic product either on a national basis or from a global perspective. Fujiwara et al. (2010) considered the production network formed by a million firms and millions of supplier-customer links, showing "in the empirical analysis scale-free degree distribution, disassortativity, correlation of degree to firm-size, and community structure having sectoral and regional modules". Moreover, Lee et al. considered the influence of the global economic network

topology to the spreading of economic crises, showing, by means of network dynamics, that its connectivity in the global network conditions the role of the nation in the crisis propagation together with its macroeconomic indicators (Lee et al. 2011).

The problem of measuring the degree of globalization of the economy is also an outstanding question. One method is to search for a measure of the clustering features through the notion of economic distance. According to Miskiewicz and Ausloos, it is possible to develop a distance and graph analysis for doing so (Miskiewicz and Ausloos 2006). This has been performed on the GDP of G7 countries over 1950–2003. In fact, several (4) different distance functions can be used and the results compared. Moreover, the graph method takes two forms in Miskiewicz and Ausloos (2006), i.e. (a) a unidirectional or (b) a bidirectional chain. In brief, the (linear) network is allowed to grow, accumulating distances, in one or two directions.

Defining the percolation transition threshold, as the distance value at which all countries are connected to the network, it has been found that the correlations between GDP yearly fluctuations (Miskiewicz and Ausloos 2006; Gligor and Ausloos 2007) achieve their highest value in 1990. Hence, the globalization of the world economy was seen to disappear as early as 2005 in publications, and is so confirmed nowadays by economists.

Macroeconomic indicators other than GDP correlations can be used for such globalization/infinite cluster search. In Gligor and Ausloos (2007), 11 of them were investigated for the 15 original EU countries,—the data taken between 1995 and 2004. Moreover, besides the Correlation Matrix Eigensystem Analysis, a Bipartite Factor Graph Analysis is of interest for confirming the existence of stable "economic" communities. It has been interestingly found that strongly correlated countries, with respect to these 11 macroeconomic indicators fluctuations, can be partitioned into clusters, mainly based on geographic grounds.

The Moving Average Minimal Length Path algorithm (Gligor and Ausloos 2008c) has allowed a decoupling of the fluctuations, whence a Hamiltonian representation can be formulated, given by a factor graph. Practically, that means that the Hamiltonian-Liouville machinery could be used thereafter. In fact, the Hamiltonian plays the role of a cost function.

It is somehow evident that markets can not be instantaneously correlated. In Ausloos and Lambiotte (2007a), forward and backward correlations, and distances, between GDP fluctuations were calculated for 23 developed countries. In this study, the network links were not only weighted but were also directed due to the arrow of time. Filtering the time-delayed correlations by successively removing the least correlated links, an evolution of the 23 countries network can be visualized. In so doing, this percolation idea-based method reveals the emergence of connections, but interestingly also of leaders and followers.

It is also relevant to know what time window has to be used when averaging properties, both in micro- and macro-economic considerations. In Gligor and Ausloos (2008c), several statistical distances between countries were calculated for various moving time windows. Up to four macroeconomic indicators were investigated:

GDP, GDP/capita (GDPpc), Consumption, and Investments. In using a so called optimal one, some empirical evidence has been presented to indicate economic aspects of globalization through such indicators. The hierarchical organization of countries and their relative movement inside the hierarchy have been described.

On the practical side, there are policy implications concerning the economic clusters arising in the presence of Marshallian externalities. Relationships between trade barriers, R&D incentives and growth were identified. It is recommended that they should be accounted for designing a cluster-promotion policy (Ausloos and Lambiotte 2007a). Not only it should be admitted that fluctuations in macroeconomic indicators result from nearby node policies, but the effect of policy changes are not immediate. Therefore, some time averaging is necessary in order to find out, how long it takes before some policy affects some country economy and its neighboring or its connected countries. As in Gligor and Ausloos (2008c), an investigation of the weighted fully connected network of the 25 countries (nodes) forming the European Union in 2005 was presented (Gligor and Ausloos 2008b) to study such time parameter effects. The links were taken to be proportional to the degree of similarity between the macroeconomic fluctuations and the GDP/capita (GDPpc) annual rates of growth between 1990 and 2005, measured by the "coefficients of determination" (Gligor and Ausloos 2008b). It has been found that the effect of the time window size for averaging and finding some robust correlations was a 7-years time window; this time interval leads to coherence in the data analysis and subsequent interpretations. A calculation of the "overlap index" (Gligor and Ausloos 2008b) reveals the emergence and stability of a "hierarchy" among EU countries (Gligor and Ausloos 2008b,c).

The Cluster Variation Method for weighted bipartite networks, discussed in Sect. 2, was applied in Gligor and Ausloos (2008a). The method allows to decompose (or expand) a "Hamiltonian" through a finite number of components, serving to define variable clusters in a given (fully connected) network. In the case studied in Gligor and Ausloos (2008a), the network was built from data representing correlations between four macro-economic features: Gross Domestic Product (GDP), Final Consumption Expenditure (FCE), Gross Capital Formation (GCF) and Net Exports. Two interesting features were deduced from such an analysis: (a) the *minimal entropy clustering scheme* is obtained from a coupling *necessarily* including GDP and FCE; (b) the *maximum entropy* corresponds to a cluster which does *not explicitly include* the GDP.

A new methodological framework for investigating macroeconomic time series has been introduced in Redelico et al. (2009), based on a non-linear correlation coefficient. Features related directly to the Latin American, rather than EU, countries have been described, i.e. those Latin American countries where Spanish and Portuguese languages prevail. Again, clusters, in the networks, and emergence of a "hierarchy" have been identified through the "Average Overlap Index" (Gligor and Ausloos 2008b) hierarchy scheme.

A principal component analysis has further been applied in order to observe and corroborate whether a country clustering structure truly exists (Redelico et al. 2009), and the results confirmed previously expected results.

4.2 Finance: Market Correlations and Concentration

In this section some literature concerning the application of complex networks to financial problems will be reviewed, without any particular attention being paid to the problem of banking risk, which is treated in another chapter of this book (Bougheas and Kirman 2015). Besides using a network structure, risk is widespread modelled through correlations between different financial instruments. However, these correlations wholly describe risk distribution functions only if the phenomena under observation are Gaussian, since they correspond to the second order moment of the joint probability distribution. Higher order models could be relevant in more general situations, giving rise to deviations from the Gaussian distribution. This notwithstanding, it is worth mentioning that in option pricing, although there is evidence of deviations from the Gaussian hypothesis (Cerqueti and Rotundo 2010a), most models used by practitioners predominantly rely on it and thus give rise to self-fulfilling prophecies (Levy et al. 2000). There is a plethora of quantitative and econometric models to analyse option pricing, while the complex-network perspective for understanding the relative distance among systems and defining clusters is much more recent. In relation to the complex-network approach it has been remarked that filtering financial data is relevant for introducing a measure of distance based on stock market indices (Mantegna 1999; Pozzi et al. 2008). Subsequent interdisciplinary studies show how to extract relevant information through methods first proposed in graph theory, namely Maximum Spanning tree and the Planar Maximally Filtered Graph also under dynamical change of data (Pozzi et al. 2008, 2013; Caldarelli et al. 2003).

Yet, the correlation matrix is not the only instrument available for understanding systemic risk due to the connections among companies. Both the cross-shareholding matrix and the board of directors describe ties among companies, that may cause cascade spreading of financial crises, and raise the question of who is controlling the controller, or having a relative size on markets sufficient to pass the critical threshold of market concentration, triggering the antitrust measures (Rotundo and D'Arcangelis 2010a,b, 2013, 2014; Grassi 2010; Croci and Grassi 2013; D'Errico et al. 2008; Rotundo 2011). The development of algorithms for detecting market leaders and related optimization problems is studied also using operation research methods (Salvemini et al. 1995), thus confirming the interdisciplinary nature of the field. Centrality measures on networks are applied for their purpose of evidencing leaders, and consider antitrust policies for the markets. Moreover, due to the globalization of trading, the default or crisis of companies in one country may spread worldwide (Garas et al. 2010, 2012; Vitali et al. 2011).

As recently claimed (Lux and Westerhoff 2009), "economic theory failed to envisage even the possibility of a financial crisis like the present one. A new foundation is needed that takes into account the interplay between heterogeneous agents." This foundation can be found in the field of complex networks, as stated by Catanzaro and Buchanan in the same issue (Catanzaro and Buchanan 2013). Several other papers in the same special issue of Nature Physics contribute different applications of complex networks to the field of finance (Caldarelli et al. 2013; Galbiati et al. 2013)

Pursuing this new foundation, a great number of results have been reported in the recent past concerning the application of complex networks formalism to the analyses of financial problems. Probably, these new topological objects have not been applied to any other field of economics to a larger extent. Sampling the work in this area we will mention the works of Onnela, Saramäki, Kaski and coworkers (Onnela et al. 2002, 2003a,b,c, 2004a,b, 2006; Saramäki et al. 2005; Onnela 2006) for an extensive application of dynamic asset trees to financial markets, as well as to clustering, communities and correlations in these systems.

Oatley et al. analyzed the political economy of global financial network using a network model (Oatley et al. 2013), and Caldarelli et al. also employed this formalism and statistical mechanics for reconstructing a financial network even from partial sets of information (Caldarelli et al. 2013). On the other hand, da Cruz et al. applied non-equilibrium statistical physics to a system of economic agents obeying the Merton–Vasicek model for current banking regulation and forming a network of trades by means of the exchange of an "economic energy" (da Cruz and Lind 2012). The authors analyzed the propagation of insolvency (i.e. the falling of an agent below a minimum capital level) in this network and were able to prove that the avalanche sizes are governed by power-law distributions whose exponents are related to the minimum capital level. Avalanches have been proved to occur also due to behavioral aspects like the blindness to small changes in the worldwide network of stock markets (Vitting Andersen et al. 2011). Finally, we will mention the work of Bonanno et al. (2004) who considered correlation-based networks of financial equities.

4.3 Tax Evasion

The network approach has also been applied to the study of tax evasion (Zaklan et al. 2008). The authors use the standard two-dimensional Ising model treated in Sect. 2 to analyze the effect of the structure of the underlying network of taxpayers on the time evolution of tax evasion in the absence of measures of control. Furthermore, it is shown that "even a minimal enforcement level may help to alleviate this problem substantially". The number of applications of the Ising model is thus augmented suggesting an enforcement mechanism to policy-makers for reducing tax evasion.

Moreover Zaklan et al. allowed tax evaders to be randomly subjected to audits, assuming that if they get caught they behave honestly during certain time (Zaklan et al. 2009). Considering different combinations of parameters, they proved that using punishment as an enforcement mechanism can effectively control tax evasion.

4.4 Business and Spreading of Innovations

Beyond the numerous applications of complex networks reviewed in the previous section, these networks have been also applied in other fields of Economics. This formalism is particularly useful for describing the introduction of innovations in markets and/or regions, and has been thoroughly used for that purpose. This methodology has also been used for analyzing business networks. Here some contributions devoted to the study of profit optimization under technological renewal are reviewed, along with dynamic models of oligopoly with R&D externalities on networks, topics on upstream/downstream R&D networks and welfare and the spreading of products in markets or co-workers networks.

Diffusion in complex social networks has been considered by several authors. López-Pintado analyzed the spreading of a given behavior in a population by considering mutual neighbor influence in a network of interacting agents by means of a simple diffusion rule (López-Pintado 2008a). At a mean-field level, she obtained a threshold for the spreading rate for propagation and persistence in populations, This threshold depends on the connectivity distribution of the underlying social network as well as on the selected diffusion rule. More recently, the same author considered the spread of free-riding behavior in social networks introducing a model for a social network with free-riding incentives, where agents are allowed to decide whether or not to contribute to the provision of a given local public good (López-Pintado 2008b). By means of equilibria analysis of the induced game, the author reported the influence of the degree distribution of the underlying network in the fraction of free-riders. Moreover, López-Pintado and Watts addressed the problem of the collective behavior of individuals facing a binary decision under the influence of their social network (López-Pintado and Watts 2008). The authors reported both the equilibrium and non-equilibrium properties of the collective dynamics and a response function under global and anonymous interactions.

Concerning business structure, the work (Semitiel-García and Noguera-Méndez 2012) is relevant, who, using network theory and social network analysis analyzed the influence of inter-industrial structures and the location of economic sectors, on the diffusion of knowledge and innovation. Specifically, they studied the structure and dynamics of the Spanish Input–output system over a 35-year period.

Business networks have also been considered by Souma et al., who categorized them into bipartite networks, showing the possibility that business networks will fall into the scale-free category (Souma et al. 2003). By means of a one-mode reduction the authors were able to approximately calculate the clustering coefficient and the averaged path length for bipartite networks. These quantities were calculated for

networks of banks and companies before/after a bank merger, and they reported quantitative evidence that banks merging increases the cliquishness of companies, and decreases the path length between two companies.

In Cerqueti and Rotundo (2010b) and Cerqueti and Rotundo (2007, 2009) the authors developed models for a set of firms producing a single commodity dealing with the optimal time for the renewal of the technology. Such models consider the aggregate outcome. Eventually, the presence of a hierarchical network organization among firms allows the leader company to propose a financial strategy, but the proposal is followed by the firms at the peripheral of the network with a certain probability only. Depending on the connection level among the companies conditions are obtained for the best strategies to optimize the profit of a district when a technological renewal takes place. The papers refer to empirical results drawn on the most large databases CENSUS and COMPUSTAT for shaping the density of companies and the studies are well suitable for the development of policies for industrial districts (Amaral et al. 1997a,b; Axtell 2001).

With respect to the use of complex networks in oligopoly analysis, in a series of papers (Bischi and Lamantia 2012a,b), Bischi and Lamantia considered the possibility of reducing the cost of knowledge gain for firms through sharing R&D as well as through investments in R&D cost-reducing activities. These researchers introduced a two-stage oligopoly game for which they analyzed the existence and stability of equilibria in a given network divided into sub-networks. In pursuing such considerations, the authors considered, in the framework of the two-stage oligopoly game, the influence of the degree of collaboration and spillovers on profits, social welfare and overall efficiency (Bischi and Lamantia 2012a). Analytical results are provided for two relevant cases performing numerical experiments and emphasizing the role of the level of connectivity (i.e. the collaboration attitude) inside networks. The effects of unintentional knowledge spillovers inside each network and between competing networks are also considered in Bischi and Lamantia (2012b).

The endogenous formation of *upstream* R&D networks have been studied in a vertically related industry and the welfare implications thereof Kesavayuth et al. (2014). The authors reported that in the situation where *upstream* firms fix prices, the complete network of firms reaches an equilibrium. In contrast, if upstream firms set quantities, a complete network arise only for sufficiently low R&D spillovers between the firms If these R&D spillovers are sufficiently high, a partial network arises. Hence, socially optimal equilibrium networks are only reached if upstream firms set prices, and the actual behavior of upstream firms must be taken into account when designing technology policy, and not only the size intra-network R&D spillovers (Kesavayuth et al. 2014). The downstream firms' incentives in a vertically organized industry have also been examined in Manasakis et al. (2014), where the authors analyzed how and when to invest in cost-reducing R&D, and to form a Research Joint Venture (RJV). The authors identified conditions for an RJV to be beneficial to society and discussed integrated innovation and competition policies.

Dal Forno and Merlone have considered a network of individuals supposing that they could propose and successfully implement their best project (Dal Forno and Merlone 2007). Important elements in the network are: (a) mutual knowledge, (b) agent coordination in choosing the project to implement, (c) the number of leaders, and (d) their location. Leaders increase the social network of other agents making possible projects otherwise impossible; at the same time, they are crucial in setting the pace of a balanced expansion of the social matrix. According to evidence, leaders are not those with the greatest number of connections. The presence of leaders provides a solution to the selection problem when there are multiple equilibria.

Finally, Pombo et al. presented evidence of the existence of imitative behaviour among family practitioners in Galicia (Spain), and they used complex network theory and the Ising model (see Sect. 2) in order to describe the entry of new drugs in the market, treating doctors as spins (nodes) in a Watts–Strogatz network (Pombo-Romero et al. 2013). Related to this, one could mention research work done on self-citations of coauthors as defining their research field flexibility, curiosity and in some sense creativity (Ausloos et al. 2008; Hellsten et al. 2006, 2007). A combination of such investigations might not only lead to new methods for detecting scientists field mobility, but also indicate pertinent features on new ideas related to the evolution of the production of new goods.

5 Regional Trade, Mobility and Development

Regional trade is another field of economics that has been extensively treated within the framework of complex network formalism. Specifically, a lot of effort has been devoted to the description of the structure (Bhattacharya et al. 2008; Rodrigue 2013; Kali and Reyes 2007; Garlaschelli and Loffredo 2005; Fronczak and Fronczak 2012; Fagiolo et al. 2008), communities (Barigozzi et al. 2011) and dynamics (Xiang et al. 2003) of the global trade network. Specifically, in the recent past Fronczak and Fronczak reported a statistical mechanics study of the international trade network showing that this network is a maximally random weighted network, and that the product of the GDP's of the trading countries is the only characterizing factor of the directed connections associated to bilateral trade volumes (Fronczak and Fronczak 2012). Moreover, Reyes et al. considered the bilateral trade data from the networks perspective, concluding that there seems to be a cyclical pattern in the regional trade agreement formation on the community structure of the world trade network (Reyes et al. 2009). From this perspective, the pattern of international integration followed by East Asian countries and its comparison with the Latin American performance has been also reported recently (Reyes et al. 2010).

In this subsection we report contributions that, although not truly devoted to network analysis, depend on the existence of a (fully connected) network.

Among the empirical and quantitative studies, Paas and Schlitte studied the regional income inequality and convergence process in the EU-25 (Paas and Schlitte 2008). Paas and Vahi considered the contribution of innovation to regional

disparities and convergence in Europe using empirical GDPpc and innovation indicators of the EU-27 NUTS2 the regions (Paas and Vahi 2012). Using principal components factor analysis, three composite indicators of regional innovation capacity were extracted, showing that ca. 60 % of variability of regional GDP per capita is associated to regional innovation performance. Regional innovations are seen to promote the increase of inter-regional differences in the short-run. Consequently, further policy interventions beyond innovation activities should be effectively implemented. In this respect see also the related work by Gligor and Ausloos, already mentioned (Gligor and Ausloos 2008a,b,c), on globalization and hierarchical structures in EU, as well as the need of considering appropriate forward and backward correlations within appropriate time intervals (Miskiewicz and Ausloos 2006; Ausloos and Lambiotte 2007a).

As can be seen in another chapter of this book by Nelson, it is common understanding that international mobility of people and workers are increasing globally (Nelson 2015). According to Paas and Halapuu, an ethnically and culturally diverse population is expected to create greater variability in the demand for goods and services as well as in the supply of labour through different skills and business cultures favoring new business activities and future economic growth (Paas and Halapuu 2012). The authors state that "although not all immigrants are well-educated and highly-skilled to provide a sufficiently innovative and creative labour force, national economic policies should create conditions that support the integration of ethnic diversity in order to create stable and peaceful environment for economic and political development". Paas and Halapuu aim at clarifying the possible determinants of peoples' attitudes towards immigrants depending on their personal characteristics, as well as attitudes towards households socio-economic stability and a country's institutions. The study's overwhelming aim is to provide empirical evidence-based reasons for policy proposals that, through integration of ethnically diverse societies, creates a favorable climate supporting economic growth. Based on the formulated aims, Paas and Halapuu used principal component factor analysis and micro-econometric methods data from the European Social Survey (ESS) fourth round database to examine the attitudes of European people towards immigrants (Paas and Halapuu 2012). These attitudes of the European people's towards immigration, which strongly constraints mobility between regions, are proven to depend on several factors such as the personal characteristics of the respondents, the attitudes towards the country's institutions and socio-economic security, and, finally, country specific conditions.

On the analytic side, the work (Vitanov and Ausloos 2012) is noteworthy. Even though the authors are not considering a network per se, they nevertheless include spatial gradients between regions in order to study issues such as: knowledge epidemics that take into account population dynamics and models describing the diffusion of ideas. This work relies on the use of the Lotka–Volterra system of equations with spatial gradients between regions with the addition of demographic input.

6 Other Social Network Models

As outlined here above, network theory is increasingly gaining acceptance in the economics community in order to understand the Economy as an evolving complex structure of widely interacting heterogeneous agents (Ausloos et al. 2014). The alternative view to that is to consider the economy as an archipelago of *homo economicus* individuals, still interacting, but on a "shorter range", underlying the neoclassical economy (Walras 1954) mainstream features. However, in Sociology and other disciplines like Ecology, Computation, etc., the network perspective has been adopted for decades.

In fact, the origin of considering a society as made of interacting entities goes back to Comte (1852, 1995), the founder of the discipline of sociology, having introduced the term as a neologism[1] in 1838, among other scientific contributions. Note that Comte had earlier used the term "social physics", but that term had been appropriated by others, e.g. Quetelet (1835, 1869). Thereafter, Boltzmann and Maxwell imagined similar concepts for describing matter and natural phenomena. Needless to say that much work has followed since.

Many examples of applications of the network approach to social sciences can be found in the literature. Reviews of this work can be found in Kirman (1992) and Stauffer (2012), in the compendium (Chakrabarti et al. 2007) and in the recent book (Galam 2012) further elaborating the work in Galam (2008). Other studies have focused on computational techniques (Stauffer 2003). In more recent times the approach has been widely accepted and acknowledged (Săvoiu and Iorga-Simăn 2012). The various methods used by physicists, applied mathematicians, economists, social scientists, differ much from each other due to the targets and methods of analysis. For instance, it is the law of large numbers which allows the application of statistical physics methods (Stauffer 2003). As far as the application of networks for social modelling is concerned, they contribute to develop tools that allow social scientists to understand how and when social factors such as peer influences, role models, or norms affect individual choices (Sousa et al. 2005; Bischi and Merlone 2010). In what follows, several applications of complex networks are reviewed all aiming at describing some type of social networks.

A large number of studies have applied complex networks to the study of systems like the Internet and the World Wide Web (WWW), and they have been extensively summarized in the reviews cited earlier. However, for an updated review focused specifically on applications the reader is referred to da Costa et al. (2011), where applications of complex networks to real-world problems and data are reported. The authors surveyed the applications of complex networks formalism in "no less than 11 areas, providing a clear indication of the impact of the field of complex networks". Moreover, the book by Vega-Redondo (2007) provides a comprehensive

[1]The word was first coined by Siey's in 1780 (Guilhaumou 2006).

coverage with applications of complex networks to labor markets, peer group effects, trust and trade, and research and development.

Social networks are known to be organized into densely connected communities, with a high degree of the clustering, and being highly assortative. The formation of complex networks has been reported from the experimental perspective by Bernasconi et al. (2010) using non-cooperative games of network formation based of the Bala and Goyal type (Bala and Goyal 2000). On their side, Toivonen et al. reported a realistic model for an undirected growing network for its use in sociodynamic phenomena (Toivonen et al. 2006). On the other hand, Boguñá et al. used an abstraction of the concept of social distance to define a class of models of social network formation (Boguñá et al. 2004). The evolution of structure within large online social networks is examined in Kumar et al. (2010), with specific attention focused on the Flickr and Yahoo's social network, showing their segmentation, and providing a detailed characterization of the structure and evolution of their different regions as well as a simple model of network growth capable of mimicking this structure (Kumar et al. 2010). Moreover, Palla et al. quantified social group evolution by means of an algorithm based on clique percolation for the time dependence of overlapping communities on a large scale (Palla et al. 2007). Finally, it is noteworthy that the problem of the determination of the community structure in the presence of unobserved structures among the nodes—a rather common situation in social and economic networks—has been addressed by Copic et al., who axiomatically introduced a maximum-likelihood-based method of detecting the latent community structures from network data (Copic et al. 2009).

Opinion formation in social systems has also been a matter of great concern in network literature. Apart from some pioneering work like that of Kirman and collaborators in the 1980s using the diameter-2 Bollobás model (Kirman et al. 1986), the field has evolved only recently when a plethora of contributions have been reported. As a very recent example, Koulouris et al. reported the multi-equilibria regulation of opinion formation dynamics (Koulouris et al. 2013).

In opinion formation models the single vote of an individual can be influenced and can change, but when the final target is the aggregate, the sampling of many randomly selected people can give a reasonable impression for an upcoming election (Stauffer 2012). In economics agent-based models have been used for the analysis of the aggregate behaviour of a large number of individuals as model with heterogeneous agents are gaining a more prominent role relative to those with a representative agent (Kirman 1992).

Network theory has been also applied to other social networks of interest like opinion formation, social entrepreneurship, etc. Dal Forno and Merlone adapted the notion of density of a graph to multiple projects and non-dichotomous networks. An appropriate visualization procedure has been implemented in Dal Forno and Merlone (2008). Social entrepreneurship effects on the emergence of cooperation in networks have been examined in Dal Forno and Merlone (2009), where differences between social entrepreneurs and leaders are analyzed and where the network of interactions may allow for the emergence of cooperative projects. The model reported by the authors consists on two coupled networks standing for knowledge

and cooperation among individuals respectively. Any member of the community can be a social entrepreneur. On the basis of this theoretical framework, the authors prove that a moderate level of social entrepreneurship is enough for providing a certain coordination on larger projects, suggesting that a moderate level of social entrepreneurship would be sufficient.

Lambiotte and Ausloos analyzed the coexistence of opposite opinions in a network with communities (Lambiotte and Ausloos 2007a). Applying the majority rule to a topology with two coupled random networks, they reproduced the modular structure observed in social networks. The authors analytically calculated the asymptotic behavior of the model deriving a phase diagram that depends on the frequency of random opinion flips and on the inter-connectivity between the two communities. Three regimes were shown to take place: a disordered regime, where no collective phenomena takes place; a symmetric regime, where the nodes in both communities reach the same average opinion; and an asymmetric regime, where the nodes in each community reach an opposite average opinion, registering discontinuous transitions from the asymmetric regime to the symmetric regime.

In this same model, Lambiotte et al. have shown that a transition takes place at a value of the interconnectivity parameter, above which only symmetric solutions prevail (Lambiotte et al. 2007). Thus, both communities agree with each other and reach consensus. Below this value, the communities can reach opposite opinions resulting in an asymmetric state. They explicitly analyzed the importance of the interface between the subnetworks.

Finally Lambiotte and Ausloos studied collaborative tagging as a tripartite network, analyzing online collaborative communities described by *tripartite* networks whose nodes are persons, items and tags (Lambiotte and Ausloos 2006a). Using projection methods they uncovered several structures of the networks, from communities of users to genre families.

Finally, two economico-sociological studies are outlined. One pertains to the interaction of small world networks of biased communities, like the neocreationists vs. the evolution defenders. For this analysis, the networks are considered to be directed but with unweighted links (Ross and Ausloos 2009; Rotundo and Ausloos 2010). The other study mentioned above pertains to bipartite networks made of music listeners downloading some music work from the web (Lambiotte and Ausloos 2005b, 2006c).

7 Suggestions for Future Research

The application of the theory of complex networks, alone or in combination with other theoretical developments of statistical mechanics, can lead to very interesting results in several areas of economics in particular, and of sociophysics in general. Specifically, the elaboration of a model for the fluxes of entrepreneurs, trade and workers between the European regions undoubtedly demands the usage of complex networks together with non linear model for the dynamics of the agents. This

treatment should generalize early regional science contributions, which are typically based on flow equations theories of directed diffusion (see for e.g. those in Hotelling 1929; Beckman 1952; ten Raa 1986; Puu 1982 where goods and migration fluxes are governed by conventional systems of diffusion-like partial differential equations). Moreover, this treatment could be very well complemented by a nonlinear model of production and consumption cycles, in the line of Meadows Dynamics of Commodity Production Cycles (Meadows 1970). This could be done in analogy to what has been done for biological oscillators (see for example Goodwin's model of enzyme production in Goodwin 1965; Murray 2002 for a classical review on these kind of models). The introduction of spatial inhomogeneities in these models could also provide a new research path for economic geography.

A list of other potentially fruitful research avenues is provided below.

Innovation and Renewal of Technology It could be useful to apply complex networks and agent-based models to the analysis of the spreading of technological renewal, R&D incentives and growth, fiscal (regional) rules (Heppenstall et al. 2012) usefulness of data analyses through rescale range analysis methods, principal component analysis. Moreover, this framework can also be applied to transportation, migration, growing and diversifying nodes of networks; merging and controls of agents; or tendency toward monopoly through lobbying.

Regional Trade and Development The network approach can also be applied for introducing relationships like trade barriers, community detection, clusters, hierarchies; policy implications concerning the economic (regional) clusters arising in the presence of Marshallian and other externalities, etc.

Development of Database and Data Mining "I would not have thought that the spread (IT/DE) was going to rise again" [IT politician, summer 2012]. Economic and financial theories need to be tested on real world markets. The complexity and large amount of data makes impossible autonomous data collection at the individual level. Moreover, data providers as the Bureau van Dijk or Bloomberg do implement only certain type of data retrieval, and the work that researchers have to do autonomously delays the production of results, and the detection of information that can be used as input for more complex models. Moreover, the increasing costs of subscriptions to data providers, in conjunction with the progressive decrease of national funding, suggest that the development of synergies with data providers is timely.

Statistical Mechanics Approaches In future works, phase transitions, coupling between magnetic and cristallographic transitions, thermodynamic (through the notion of cost function) vs. geometric (percolation) could be analyzed, including instabilities in necessarily non equilibrium structures (log-periodic oscillations). Moreover, going beyond Ising model (Blume-Emery-Griffiths, and the forgotten ferroelectric models) should be considered, as well as community detection, forward and backward correlations in networks with weighted and directed links (danger of difficulty in interpreting complex eigenvalues of adjacency matrix), network structure construction and evolution, etc.

Economic and Financial Networks and Risk " When Belgium sneezes, the world catches a cold" [http://phys.org/news/2010-11-belgium-world-cold.html] Globalization of economic and financial markets, corporate ownership networks, international trade networks, as well as phenomena like tunnelling, cross-ownership, and boards interlock, change dramatically the profile of financial and economic risks pointing out the relevance of the network structure and are topics to be considered in the future. Understanding such phenomena both at the micro and macro level may help the development of policies also at the local level with potential benefits for regional trade and development.

8 Conclusions

The range of applications of complex networks formalism is expanding at a fabulous rate, and has been adopted almost in every field of knowledge having to deal with heterogeneous interacting agents and their emergent phenomena. This is the case, particularly, of economics, and even more specifically of economic geography. The present report gathers contributions from different fields and approaches under the common theme of complex networks analysis in socioeconomic models. The statistical mechanics of complex networks have been reviewed together with some computational aspects related to their description. Models specifically developed for examining topics in various areas of economics and finance, such as, for example, regional trade and development have been the object of specific attention, together with contributions devoted to the application of complex networks analysis to social networks in the broad sense.

Acknowledgements This work has been performed in the framework of COST Action IS1104 "The EU in the new economic complex geography: models, tools and policy evaluation".

References

Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*, 47–97.

Almeida-Neto, M., Guimarães, P., Guimarães, P. R., Jr., Loyola, R. D., & Ulrich, W. (2008). A consistent metric for nestedness analysis in ecological systems: Reconciling concept and measurement. *Oikos, 117*, 1227–1239.

Amaral, L., Buldyrev, S. V., Havlin, S., Leschhorn, H., Maass, P., Salinger, M. A., Stanley, H. E., & Stanley, M. H. R. (1997a). Scaling behavior in economics: I. Empirical results for company growth. *Journal de Physique I, 7*, 621–633.

Amaral, L., Buldyrev, S. V., Havlin, S., Leschhorn, H., Maass, P., Salinger, M. A., Stanley, H. E., & Stanley, M. H. R. (1997b). Scaling behavior in economics: II. Modeling of company growth. *Journal de Physique I, 7*, 635–650.

Araujo, A. I. L., Corso, G., Almeida, A. M., & Lewinsohn, T. M. (2010). An analytic approach to the measurement of nestedness in bipartite networks. *Physica A, 389*, 1405–1411.

Arthur W. B. (2006). Out-of-equilibrium economics and agent-based modelling. *Handbook of Computational Economics, 2*, 1551–1564.

Ausloos, M., & Lambiotte, R. (2007a). Clusters or networks of economies? A macroeconomy study through GDP fluctuation correlations. *Physica A, 382*, 16–21.

Ausloos, M., & Lambiotte, R. (2007b). Drastic events make evolving networks. *European Physical Journal B, 57*, 89–94.

Ausloos, M., Lambiotte, R., Scharnhorst, A., & Hellsten, I. (2008). Andrzej Pekalski networks of scientific interests with internal degrees of freedom through self-citation analysis. *International Journal of Modern Physics C, 19*, 371–384.

Ausloos, M., Dawid, H., & Merlone, U. (2014). Spatial interactions in agent-based models. In P. Commendatore, S. Kayam, & I. Kubin (Eds.), *Complexity and geographical economics: Topics and tools*. Heidelberg: Springer.

Avrutin, V., Levi, P., Schanz, M., Fundinger, D., & Osipenko, G. S. (2006). Growing network with *j* -redirection. *International Journal of Bifurcation and Chaos, 16*, 3451–3496.

Axtell, R. L. (2001). Zipf distribution of U.S. firm sizes. *Science, 293*, 1818–1820.

Bala, V., & Goyal, S. (2000). A noncooperative model of network formation. *Econometrica, 68*, 1181–1229.

Barabási, A. L. (2003). *Linked how everything is connected to everything else and what it means for business, science, and everyday life*. New York: Plume Books.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*, 509–512.

Barigozzi, M., Fagiolo, G., & Mangioni, G. (2011). Community structure in the multi-network of international trade complex networks. *Communications in Computer and Information Science, 116*, 163–175.

Barkley Rosser, J., Jr. (1999). On the complexities of complex economic dynamics. *The Journal of Economic Perspectives, 13*, 169–192.

Barrat, A., & Weigt, M. (2000). On the properties of small-world network models. *European Physical Journal B, 13*, 547–560.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences, 101*, 3747–3752.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*.

Batagelj, V., & Mrva, A. (2003). Pajek-analysis and visualization of large networks. In M. Jünger & P. Mutzel (Eds.), *Graph drawing software* (pp. 77–103). Berlin: Springer.

Beckman, M. (1952). A continuous model of transportations. *Econometrica, 20*, 643–660.

Bernasconi, M., & Galizzi, M. (2010). Network formation in repeated interactions: Experimental evidence on dynamic behaviour. *Mind Society, 9*, 193–228.

Bertoni, F., & Randone, P. A. (2006). *The small-world of Italian finance: Ownership interconnections and board interlocks amongst Italian listed companies*. Technical Report Politecnico di Milano.

Bhattacharya, K., Mukherjee, G., Saramaki, J., Kaski, K., & Manna, S. S. (2008). The international trade network. In *Econophysics of markets and business networks, new economic windows series* (pp. 139–147). Berlin: Springer.

Bianconi, G. (2002). Mean-field solution of the Ising model on a Barabási-Albert network. *Physics Letters A, 303*, 166–168.

Bischi, G. I., & Lamantia, F. (2012a). A dynamic model of oligopoly with R&D externalities along networks. Part I. *Mathematics and Computers in Simulation, 84*, 51–65.

Bischi, G. I., & Lamantia, F. (2012b). A dynamic model of oligopoly with R&D externalities along networks. Part II. *Mathematics and Computers in Simulation, 84*, 66–82.

Bischi, G. I., & Merlone, U. (2010). Global dynamics in adaptive models of collective choice with social influence. In G. Naldi (Ed.), *Mathematical modelling of collective behavior in socio−economic and life sciences* (Vol. 223–244). Berlin: Springer.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D.U. (2006). Complex networks: Structure and dynamics. *Physics Reports, 424*, 175–308.

Boguñá, M., & Pastor-Satorras, R. (2002). Epidemic spreading in correlated complex networks. *Physical Review E, 66*, 047104.

Boguñá, M., Pastor-Satorras, R., Díaz-Guilera, A., & Arenas, A. (2004). Models of social networks based on social distance attachment. *Physical Review E, 70*, 056122.

Bonanno, G., Caldarelli, G., Lillo, F., Micciché, S., Vandewalle, N., & Mantegna, R. N. (2004). Networks of equities in financial markets. *European Physical Journal B, 38*, 363–371.

Bougheas, S., & Kirman, A. (2015). Complex financial networks and systemic risk: A review. In P. Commendatore, S. Kayam, & I. Kubin (Eds.), *Complexity and geographical economics: Topics and tools*. Heidelberg: Springer

Caldarelli, G., Lillo, F., & Mantegna, R. N. (2003). Topology of correlation-based minimal spanning trees in real and model markets. *Physical Review E, 68*, 046130.

Caldarelli, G., Battiston, S., Garlaschelli, D., & Catanzaro, M. (2004). Emergence of complexity in financial networks. *Lecture Notes in Physics: Complex Networks, 650*, 399–423.

Caldarelli, G., Chessa, A., Gabrielli, A., Pammolli, F., & Puliga, M. (2013). Reconstructing a credit network. *Nature Physics, 9*, 119–197.

Catanzaro, M., & Buchanan, M. (2013). Network opportunity. *Nature Physics, 9*, 121–122.

Cayley, J. (1889). A theorem on trees. *The Quarterly Journal of Mathematics, 23*, 376–378.

Cerqueti, R., & Rotundo, G. (2007). Productivity and costs for firms in presence of technology renewal processes. *International Transactions in Operational Research, 14*, 521–534.

Cerqueti, R., & Rotundo, G. (2009). Companies' decisions for profit maximization: A structural model. *Applied Mathematical Sciences, 3*, 1327–1340.

Cerqueti, R., & Rotundo, G. (2010a). Options with underlying asset driven by a fractional brownian motion: Crossing barriers estimates. *New Mathematics and Natural Computation, 6*, 109–118.

Cerqueti, R., & Rotundo, G. (2010b). Firms clustering in presence of technological renewal processes. In T. Puu, & A. Panchuk (Eds.), *Nonlinear economic dynamics*. New York: Nova Science Publishers.

Chakrabarti, B. K., Chakraborti, A., & Chatterjee, A. (2007). *Econophysics and sociophysics; Trends and perspectives*. Weinheim: Wiley.

Colander, D., Holt, R., & Barkley Rosser J., Jr., (2004). The changing face of mainstream economics. *Review of Political Economy, 16*, 485–499.

Comte, A. (1852). *Cour de philosophie positive*. Paris: Borrani et Droz.

Comte, A. (1995). *Leçons sur la sociologie: Cour de philosophie positive: leçons 47 à 51*. Paris: Juliette Grange Flammarion.

Copic, J., Jackson, M. O., & Kirman, A. (2009). Identifying community structures from network data via maximum likelihood methods. *The B.E. Journal of Theoretical Economics, 9*, 1–40.

da Costa, L. F., Rodrigues, F. A., Travieso, G., & Villas Boas, P. R. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics, 56*, 167–242.

da Costa, L. F., Oliveira, O. N., Jr., Travieso, G., Rodrigues, F. A., Ribeiro Villas Boas, P., Antiqueira, L., Palhares Viana, M., & Correa Rocha, L. E. (2011). *Advances in Physics, 60*, 329–412.

Cotilla-Sanchez, E., Hines, P. D. H., Barrows, C., & Blumsack, S. (2012). Comparing the topological and electrical structure of the North American electric power infrastructure. *IEEE Systems Journal, 6*, 616–626.

Croci, E., & Grassi, R. (2013). The economic effect of interlocking directorates in Italy: New evidence using centrality measures. *Computational and Mathematical Organization Theory, 20*, 89–112.

da Cruz, J. P., & Lind, P. G. (2012). The dynamics of financial stability in complex networks. *European Physical Journal B, 85*, 256–265.

Dal Forno, A., & Merlone, U. (2007). The evolution of coworkers networks: An experimental and computational approach. In B. Edmonds, C. H. Iglesias, & K. G. Troitzsch (Eds.), *Social simulation: Technologies, advances and new discoveries* (pp. 280–293). Hershey (PA): Information Science Reference.

Dal Forno, A., & Merlone, U. (2008). Network dynamics when selecting work team member. In A. K. Naimzada, S. Stefani, & A. Torriero (Eds.), *Networks, topology and dynamics theory and applications to economics and social systems. Lecture notes in economics and mathematical systems* (Vol. 613, pp. 229–240). Berlin: Springer.

Dal Forno, A., & Merlone, U. (2009). Social entrepreneurship effects on the emergence of cooperation in networks. *Emergence: Complexity and Organization, 11*, 48–58.

D'Errico, M., Grassi, R., Stefani, S., & Torriero, A. (2008). Shareholding networks and centrality: an application to the Italian financial market. In A. Naimzada, S. Stefani, & A. Torriero (Eds.), *Network, topology and dynamics. Theory and applications to economics and social systems* (pp. 215–228). Berlin: Springer.

Djikstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik, 1*, 269–271.

Dorogovtsev, S. N., & Mendes, J. F. F. (2003). *Evolution of networks - from biological nets to the internet and WWW*. Oxford: Oxford University Press.

Erdös, P., & Rényi, A. (1959). On random graphs I. *Publications Mathematicae, 6*, 290–297.

Euler, L. (1736). Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae, 8*, 128–140.

Fagiolo, G., Reyes, J., & Schiavo, S. (2008). On the topological properties of the world trade web: A weighted network analysis. *Physica A, 387*, 3868–3873.

Foster, J. (2005). From simplistic to complex systems in economics. *Cambridge Journal of Economics, 29*, 873–892.

Friedman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry, 40*, 35–41.

Fronczak, A., & Fronczak, P. (2012). Statistical mechanics of the international trade network. *Physical Review E, 85*, 056113.

Fujiwara, Y., & Aoyama, H. (2010). Large-scale structure of a nation-wide production network. *European Physical Journal B, 77*, 565–580.

Galam, S. (2008). Sociophysics: A review of Galam models. *International Journal of Modern Physics C, 19*, 409–440.

Galam, S. (2012). *What is sociophysics about?* Berlin: Springer.

Galbiati, M., Battiston, S., & Delpini, D. (2013). The power to control. *Nature Physics, 9*, 126–128.

Garas, A., Argyrakis, P., Rozenblat, C., Tomassini, M., & Havlin, S. (2010). Worldwide spreading of economic crisis. *New Journal of Physics, 12*, 113043.

Garas, A., Schweitzer, F., & Havlin, S. (2012). A k-shell decomposition method for weighted networks. *New Journal of Physics, 14*, 083030.

Garlaschelli, D., & Loffredo, M. I. (2005). Structure and evolution of the world trade network. *Physica A, 355*, 138–144.

Gilbert, E. N. (1959). Random graphs. *Annals of Mathematical Statistics, 4*, 1141–1144.

Gitterman, M. (2000). Small-world phenomena in physics: The Ising model. *Journal of Physics A, 33*, 8373–8381.

Gligor, M., & Ausloos, M. (2007). Cluster structure of EU-15 countries derived from the correlation matrix analysis of macroeconomic index fluctuations. *European Physical Journal B, 57*, 139–146.

Gligor, M., & Ausloos, M. (2008a). Cluster expansion method for evolving weighted networks having vector-like nodes. *Acta Physica Polonica A, 114*, 491–499.

Gligor, M., & Ausloos, M. (2008b). Clusters in weighted macroeconomic networks: The EU case. Introducing the overlapping index of gdp/capita fluctuation correlations. *European Physical Journal B, 63*, 533–539.

Gligor, M., & Ausloos, M. (2008c). Convergence and cluster structures in EU area according to fluctuations in macroeconomic indices. *Journal of Economic Integration, 23*, 297–330.

Goodwin, B. C. (1965). Oscillatory behaviour in enzymatic control processes. *Advances in Enzyme Regulation, 3*, 425–438.

Grassi, R. (2010). Vertex centrality as a measure of information flow in Italian corporate board networks. *Physica A, 289*, 2455–2464.

Guilhaumou, J. (2006). Sieyés et le non-dit de la sociologie: du mot à la chose. *Revue d'histoire des sciences humaines, Naissance de la science sociale (1750–1850), 15*, 117–134.

Hellsten, I., Lambiotte, R., Scharnhorst, A., & Ausloos, M. (2006). A journey through the landscape of physics and beyond ─ the self-citation patterns of Werner Ebeling. *Scientometrics, 72*, 469–486.

Hellsten, I., Lambiotte, R., Scharnhorst, A., & Ausloos, M. (2007). Self-citations, co─authorships and keywords: A new method for detecting scientists' field mobility? *Scientometrics, 72*, 469–486.

Heppenstall, A. J., Crooks, A. T., See, L. M., & Batty, M. (Eds.). (2012). *Agent-based models of geographical systems*. Dordretch: Springer.

Herrero, C. P. (2002). Ising model in small-world networks. *Physical Review E, 65*, 066110.

Hotelling, H. (1929). Stability in competition. *The Economic Journal, 39*, 41–57.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik, 31*, 253–258.

Jackson, M. O. (2011). An overview of social networks and economic applications. In J. Benhabib, A. Bisin, & M. O. Jackson (Eds.), *The handbook of social economics*. Amsterdam: North Holland Press.

Kali, R., & Reyes, J. (2007). The architecture of globalization: A network approach to international economic integration. *Journal of International Business Studies, 28*, 595–620.

Kesavayuth, D., Manasakis, C., & Zikos, V. (2014). *Venture with upstream market power*. Working Paper.

Kirman, A. (1992). Whom or what does the representative individual represent? *The Journal of Economic Perspectives, 6*, 117–136.

Kirman, A. (1997). The economy as an evolving network. *Journal of Evolutionary Economics, 7*, 339–353.

Kirman, A., Oddou, C., & Weber, S. (1986). Stochastic communication and coalition formation. *Econometrica, 54*, 129–138.

Koulouris, A., Katerelos, I., & Tsekeris, T. (2013). Multi-equilibria regulation agent─based model of opinion dynamics in social networks. *Interdisciplinary Description of Complex Systems, 11*, 51–70.

Krapivsky, P. L., Redner, S., & Leyvraz, F. (2000). Connectivity of growing random networks. *Physical Review Letters, 85*, 4629–4632.

Kumar, R., Novak, J., & Tomkins, A. (2010). Structure and evolution of online social networks. In P. S. Yu, et al. (Eds.), *Link mining: Models, algorithms, and applications* (pp. 337–357). New York: Springer.

Lambiotte, R., & Ausloos, M. (2005a). n-body decomposition of bipartite networks. *Physical Review E, 72*, 066117.

Lambiotte, R., & Ausloos, M. (2005b). Uncovering collective listening habits and music genres in bipartite networks. *Physical Review E, 72*, 066107.

Lambiotte R., & Ausloos M. (2006a). Collaborative tagging as a tripartite network. *Lecture Notes in Computer Science, 3993*(III), 1114–1117.

Lambiotte, R., & Ausloos, M. (2006b). Modelling the evolution of coupled networks. In *First World Congress on Social Simulation e-Proceedings* (Vol. 1, pp. 375–381).

Lambiotte, R., & Ausloos, M. (2006c). On the genrefication of music: A percolation approach. *European Physical Journal B, 50*, 183–188.

Lambiotte, R., & Ausloos, M. (2007a). Coexistence of opposite opinions in a network with communities. *Journal of Statistical Mechanics, 8*, P08026.

Lambiotte, R., & Ausloos, M. (2007b). Growing network with j-redirection. *Europhysics Letters, 77*, 58002.

Lambiotte, R., Ausloos, M., & Holyst, J. A. (2007). Majority model on a network with communities. *Physical Review E, 75*, 030101.

LeBellac, M. (1992). *Quantum and statistical field theory*. New York: Oxford University Press.

Lee, K. M., Yang, J. S., Kim, G., Lee, J., Goh, K. I., & Kim, I. M. (2011). Impact of the topology of global macroeconomic network on the spreading of economic crises. *PLoS ONE, 6*, e18443. doi:10.1371/journal.pone.0018443.

Levy, H., Levy, M., & Solomon, S. (2000). *Microscopic simulation of financial markets: From investor behavior to market phenomena*. Orlando: Academic.
López-Pintado, D. (2008a). Diffusion in complex social networks. *Games and Economic Behavior, 62*, 573–590.
López-Pintado, D. (2008b). The spread of free-riding behavior in a social network. *Eastern Economic Journal, 34*, 464–479.
López-Pintado, D., & Watts, D. J. (2008). Social influence, binary decisions and collective dynamics. *Rationality and Society, 20*, 399–443.
Lux, T., & Westerhoff, F. (2009). Economic crisis. *Nature Physics, 5*, 2–3.
Manasakis, C., Petrakis, E., & Zikos, V. (2014). *Downstream research joint venture with upstream market power*. Working paper.
Mantegna, R. N. (1999). Hierarchical structure in financial markets. *European Physical Journal B, 11*, 193–197.
Martin, R., & Sunley, P. (2007). Complexity thinking and evolutionary economic geography. *Journal of Economic Geography, 7*, 573–601.
Mattis, D. C. (1976). Solvable spin systems with random interaction. *Physics Letters, 56A*, 421–422.
Meadows, D. L. (1970). *Dynamics of commodity production cycles*. Cambridge (MA): Wright-Allen Press.
Milgram, S. (1967). The small-world problem. *Psychology Today, 1*, 60–67.
Miskiewicz, J., & Ausloos, M. (2006). G7 country Gross Domestic Product (GDP) time correlations. A graph network analysis. In H. Takayasu (Ed.), *Practical fruits of econophysics*. Berlin: Springer.
Molloy, M., & Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms, 6*, 61–179.
Molloy, M., & Reed, B. (1998). The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability and Computing, 7*, 295–305.
Murray, J. D. (2002). *Mathematical biology I. An introduction*, 3rd edn. Berlin: Springer.
Namatame, A., Kaizouji, T., & Aruka, Y. (Eds.). (2006). *The complex networks of economic interactions*. Berlin: Springer.
Nelson, D. (2015). Migration and networks. In P. Commendatore, S. Kayam, & I. Kubin (Eds.), *Complexity and geographical economics: Topics and tools*. Heidelberg: Springer.
Newman, M. (2003). The structure and function of complex networks. *SIAM Reviews, 45*, 167–256.
Newman, M., Moore, C., & Watts, D. J. (2000). Mean-field solution of the small-world network model. *Physical Review Letters, 84*, 3201–3204.
Newman, M., Watts, D., & Barabási, A. L. (2006). *The structure and dynamics of networks*. Princeton (NJ): Princeton University Press.
Newman, M. E. J. (2002a). Assortative mixing in networks. *Physical Review Letters, 89*, 208701.
Newman, M. E. J. (2002b). The structure and function of networks. *Computer Physics Communications, 147*, 40–45.
Newman, M. E. J., Strogatz, S. H., & Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E, 64*, 026118.
Oatley, T., Winecoff, W. K., Pennock, A., & Danzman, S. B. (2013). The political economy of global finance: A network model. *Perspectives on Politics, 1*, 133–153.
Onnela, J. P. (2006). *Complex networks in the study if financial and social systems*. Ph.D. Thesis. http://jponnela.com/web$_$documents/t2.pdf
Onnela, J. P., Chakraborti, A., Kaski, K., & Kertész, J. (2002). Dynamic asset trees and portfolio analysis. *European Physical Journal B, 3*, 285–288.
Onnela, J. P., Chakraborti, A., Kaski, K., & Kertész, J. (2003a). Dynamic asset trees and black monday. *Physica A, 324*, 247–252.
Onnela, J. P., Chakraborti, A., Kaski, K., Kertész, J., & Kanto, A. (2003b). Asset trees and asset graphs in financial markets. *Physica Scipta, 106*, 48–54.
Onnela, J. P., Chakraborti, A., Kaski, K., Kertész, J., & Kanto, A. (2003c). Dynamics of market correlations: Taxonomy and portfolio analysis. *Physical Review E, 68*, 056110.

Onnela, J. P., Kaski, K., & Kertész, J. (2004a). Clustering and information in correlation based financial networks. *European Physical Journal B, 38*, 353–362.

Onnela, J. P., Saramäki, J., Kertész, J., & Kaski, K. (2004b). Intensity and coherence of motifs in weighted complex networks. *Physical Review E, 71*, 065103.

Onnela, J. P., Saramäki, J., Kaski, K., & Kertész, J. (2006). Financial market- a network perspective. In H. Takayasu (Ed.), *Practical fruits of econophysics. Nikkei econophysics III proceedings* (pp. 302–306). Tokyo: Springer.

Onsager, L. (1944). Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Physics Review, 65*, 117–149.

Paas, T., & Halapuu, V. (2012). *Attitudes towards immigrants and the integration of ethnically diverse societies*. Norface Migration Discussion Paper No 2012–23.

Paas, T., & Schlitte, F. (2008). Regional income inequality and convergence process in the EU-25. *Scienze Regionali: Italian Journal of Regional Science, 7*, 29–49.

Paas, T., & Vahi, T. (2012). *Regional disparities and innovations in Europe*. http://ideas.repec.org/p/wiw/wiwrsa/ersa12p80.html

Palla, G., Barabási, A. L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature, 446*, 664–667.

Pastor-Satorras, R., Rubi, M., & Díaz-Guilera, A. (Eds.). (2003). *Statistical mechanics of complex networks*. Berlin: Springer.

Pekalski, A. (2001). Ising model on a small world network. *Physical Review E, 64*, 057104.

Pissanetzky, S. (1984). *Sparse matrix technology*. New York: Academic.

Pombo-Romero, J., Varela, L. M., & Ricoy, C. (2013). Diffusion of innovations in social interaction systems. An agent-based model for the introduction of new drugs in markets. *The European Journal of Health Economics, 14*, 443–455.

Pozzi, F., Aste, T., Rotundo, G., & Matteo, T. D. (2008). Dynamical correlations in financial systems. In *Complex systems II. Proceedings of the SPIE, The International Society for Optical Engineering*, *6802*, 68021E.

Pozzi, F., Matteo, T. D., & Aste, T. (2013). Spread of risk across financial markets: Better to invest in the peripheries. *Scientific Reports, 3*, 1665.

Puu, T. (1982). Outline of a trade cycle model in continuous space and time. *Geographical Analysis, 14*, 1–9.

Quetelet, A. (1835). *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale*. Paris: Bachelier.

Quetelet, A. (1869). *Physique sociale, ou essai sur le développement des facultés de l'homme*. Paris: Muquard.

ten Raa, T. (1986). The initial value problem for the trade cycle in Euclidean space. *Regional Science and Urban Economics, 16*, 527–546.

Redelico, F. O., Proto, A. N., & Ausloos, M. (2009). Hierarchical structures in the gross domestic product per capita fluctuation in Latin American countries. *Physica A, 388*, 3527–3535.

Reyes, J., Schiavo, S., & Fagiolo, G. (2010). Using complex networks analysis to assess the evolution of international economic integration: The cases of East Asia and Latin America. *The Journal of International Trade and Economic Development, 19*, 215–239.

Reyes, J. A., Wooster, R. B., & Shirrell, S. (2009). Regional trade agreements and the pattern of trade: A networks approach. doi:10.2139/ssrn.1408784.

Rodrigue, J. P. (2013). *Transportation, globalization and international trade*. New York: Routledge.

Ross, A. G. C., & Ausloos, M. (2009). Organizational and dynamical aspects of a small network with two distinct communities: Neocreationists vs. evolution defenders. *Scientometrics, 80*, 457–472.

Rotundo, G. (2011). Centrality measures in shareholding networks. In *Use of risk analysis in computer-aided persuasion. NATO science for peace and security series* (Vol. 88, pp. 12–28). Amsterdam: IOS Press.

Rotundo, G. (2013). An investigation of computational complexity of the method of symbolic images. In A. N. Proto, M. Squillante, J. Kacpryzk (Eds.), *Advanced dynamic modeling of*

economic and social systems, Studies in computational intelligence series (Vol. 448, 109–126). Berlin: Springer.

Rotundo, G., & D'Arcangelis, A. M. (2014). Mutual funds relationship and performance analysis. *Quality & Quantity*. doi:10.1007/s11135-014-0066-z.

Rotundo, G., & Ausloos, M. (2010). Organization of networks with tagged nodes and biased links: A priori distinct communities. The case of intelligent design proponents and Darwinian evolution defenders. *Physica A, 20*, 643–660.

Rotundo, G., & D'Arcangelis, A. (2013). Network of firms: An analysis of market concentration. *Quality and Quantity*. doi:10.1007/s11135-013-9858-9.

Rotundo, G., & D'Arcangelis, A. M. (2010a). Network analysis of ownership and control structure in the Italian stock market. *Advances and Applications in Statistical Sciences, 2*, 255–273.

Rotundo, G., & D'Arcangelis, A. M. (2010b). Ownership and control in shareholding networks. *Journal of Economic Interaction and Coordination, 5*, 191–219.

Salvemini, M. T., Simeone, B., & Succi, R. (1995). Analisi del possesso integrato nei gruppi di imprese mediante grafi. *L'Industria, XVI*, 641–662.

Saramäki, J., Onnela, J. P., Kertész, J., & Kaski, K. (2005). Characterizing motifs in weighted complex networks. In J. Mendes (Ed.), *Science of complex networks. AIP conference proceedings* (Vol. 776, p. 108). New York: American Institute of Physics.

Săvoiu, G., & Iorga-Simăn, I. (2012). Sociophysics: A new science or a new domain for physicists in a modern university. In G. Săvoiu (Ed.), *Econophysics: Background and applications in economics, finance, and sociophysics*. Oxford/Waltham: Academic.

Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., & White, D. R. (2009). Economic networks: What do we know and what do we need to know? *Advances in Complex Systems, 12*, 407–422.

Semitiel-García, M., & Noguera-Méndez, P. (2012). The structure of inter-industry systems and the diffusion of innovations: The case of Spain. *Technological Forecasting and Social Change, 79*, 1548–1567.

Serrano, M. A., Krioukov, D., & Boguñá, M. (2008). Self-similarity of complex networks and hidden metric spaces. *Physical Review Letters, 100*, 078701.

Seyed-allaei, H., Bianconi, G., & Marsili, M. (2006). Scale-free networks with an exponent less than two. *Physical Review E, 73*, 046113.

Siek, J. G., Lee, L. Q., & Lumsdaine, A. (2001). *The boost graph library*. Reading (MA): Addison-Wesley.

Souma, W., Fujiwara, Y., & Aoyama, H. (2003). Growth and fluctuations of personal and company's income. *Physica A, 324*, 396–401.

Sousa, A., Malarz, K., & Galam, S. (2005). Reshuffling spins with short range interactions: When sociophysics produces physical results. *International Journal of Modern Physics C, 16*, 1507–1517.

Stauffer, D. (2003). Sociophysics— a review of recent Monte Carlo simulations. *Fractals, 11*, 313–318.

Stauffer, D. (2012). A biased review of sociophysics. *Journal of Statistical Physics, 151*, 9–20.

Tesfatsion, L. (2003). Agent-based computational economics: modelling economies as complex adaptive systems. *Information Sciences, 149*, 262–268.

Toivonen, R., Onnela, J. P., Saramäki, J., Hyvönen, J., & Kaski, K. (2006). A model for social networks. *Physica A, 371*, 851–860.

Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The anatomy of the Facebook social graph. CoRR abs/11114503.

Vega-Redondo, F. (2007). *Complex social networks*. Cambridge: Cambridge University Press.

Viana-Lopes, J., Pogorelov, G., dos Santos, J. L., & Toral, R. (2004). Exact solution of Ising model on a small-world network. *Physical Review E, 70*, 026112.

Vitali, S., Glattfelder, J. B., & Battiston, S. (2011). The network of global corporate control. *PLoS ONE, 6*, e25995.

Vitanov, N. K., & Ausloos, M. (2012). Knowledge epidemics and population dynamics models for describing idea diffusion. In A. Scharnhorst, K. Börner, & P. van den Besselaar (Eds.), *Models*

of science dynamics : Encounters between complexity theory and information sciences (pp. 69–125). Berlin: Springer.

Vitting Andersen, J., Nowak, A., Rotundo, G., Parrott, L., & Martínez, S. (2011). "Price-Quakes" shaking the world's stock exchanges. *PLoS ONE, 6*, e26472.

Walras, L. (1954). *Elements of pure economics, or the theory of social wealth*. London: Allen and Unwin.

Watts, D., & Strogatz, S. (1998). Collective dynamics of small-world networks. *Nature, 393*, 440–442.

Xiang, L., Yu, Y. J., & Guanrong, C. (2003). Complexity and synchronization of the world trade web. *Physica A, 328*, 287–296.

Yang, C. N. (1952). The spontaneous magnetization of a two dimensional Ising model. *Physical Review, 85*, 808–816.

Zaklan, G., Lima, W., & Westerhoff, F. (2008). Controlling tax evasion fluctuations. *Physica A, 387*, 5857–5861.

Zaklan, G., Westerhoff, F., & Stauffer, F. D. (2009). Analysing tax evasion dynamics via the Ising model. *Journal of Economic Interaction and Coordination, 4*, 1–14.

Part III
Industrial Interactions

Dynamics of Industrial Oligopoly Market Involving Capacity Limits and Recurrent Investment

Anastasiia Panchuk

Abstract In the current chapter we investigate an industrial oligopoly market, modelled by using CES production functions in combination with the isoelastic demand function. It is supposed that the competitors act not under constant, but eventually decaying returns, and thus, from time to time they need to renew their capital equipment, choosing its optimal amount according to the current market situation. Meanwhile, in the intervening periods the firms are subjected to capacity limits due to fixed capital stocks. As a result, the evolution of the system derived depends essentially on the number of competitors and the capital lifetime, and is also sensitive to the initial choice of individual inactivity times. In particular, the firms may merge into different groups renewing their capitals simultaneously, which leads to distinct dynamical patterns.

1 Introduction

Economic theory considers different market forms, depending on the number of competing firms (suppliers), from monopoly (one firm), over duopoly (two firms), oligopoly (a few firms), to perfect competition (a large number of firms). The consumers (demanders), on the other hand, are always supposed to be very numerous. This means that only the suppliers, in case they are few, can take strategic decisions. In the competitive case, suppliers are assumed to take price as given (unalterable by their own actions) at the equilibrium of supply and demand. A monopolist, on the other hand, knowing the demand function, i.e. the dependence of demand on price, can limit its supply in order to maximise profit, through setting a monopoly price. The case of a few competitors appears to be quite intricate, because, in addition to demand, the suppliers have to consider the outputs of their rivals as well.

A. Panchuk (✉)
Institute of Mathematics, NAS of Ukraine, 3 Tereshchenkivska Str., 01601 Kiev, Ukraine
e-mail: anastasiia.panchuk@gmail.com

P. Commendatore et al. (eds.), *Complexity and Geographical Economics*,
Dynamic Modeling and Econometrics in Economics and Finance 19,
DOI 10.1007/978-3-319-12805-4_10

The objective of this chapter is to present certain results concerning investigation of how a market is developed in case when several firms are involved. Any new industry, not depending on how large it may expand in the course of time, is originally established through a few pioneering firms, and eventually starts growing in terms of the number of competitors, thus developing competition. In what follows, we study this process through a very stylised model, based on general microeconomics. Some obvious facts are disregarded, such as evolution of technology, including efficiency differences among firms, and growing demand. Neglecting growing demand also implies that we omit the spatial elements, including transportation cost, the general decrease of which has been considered to be the main moving force behind the growth of markets. These complications may later be introduced if one wishes, but first it is good to know what happens under perfectly homogeneous and stationary conditions.

History of economic theory studying competitive markets goes back to the fundamental work of Cournot (1838). Such transitional market forms like duopoly and oligopoly are often considered to be contextually the first step from monopoly towards perfect competition. However, these intermediate cases, being investigated in a number of publications, appear to be much more complicated analytically than any of the two extreme occurrences—the case of a single dominating firm (monopoly), as well as the case of numerous small firms (perfect competition). Studies devoted to duopoly and triopoly markets already uncover presence of complex dynamical phenomena such as multistability (see, e.g., Agliari et al. 2002; Bischi et al. 2000) and homoclinic connections (see, e.g., Agliari 2006). Certain works also concern general multi-dimensional imperfect competition models including effects of partial or full cooperation (see, e.g., Kopel and Szidarovszky 2006; Szidarovszky and Okuguchi 1998). To perceive the variety of topics falling under the field of nonlinear oligopoly dynamics, see Bischi et al. (2010), Puu and Sushko (2002) and references therein.

For setting up the dynamics on the market, one needs to define how each competitor would react in the next time period to the current market situation observed. The function representing such a 'reply' is often called a *reaction function*, and may be given in different ways. One of the most simple methods [and also a traditional way to do this which comes back to Cournot (1838)] is assuming so-called *naïve expectations*. Namely, all suppliers know the output provided by the other competitors in the current time period, and they suppose that in the next time period the same amount will be produced. Thus, each firm calculates its output being produced in the time period $t + 1$ depending on the quantity produced by its rivals in the previous period t.

There is another way to define such a dynamical process which follows from the assumption that firms do not possess complete information about the global market demand function, but each of them is able to approximate its marginal profit. This leads to a so-called *gradient method*, namely, at every time period each firm adjusts its production proportionally to the obtained marginal profit approximation. This approach, although being more realistic, was anyhow applied mainly to the lower-dimensional cases (see, e.g., Ahmed et al. 2000; Angelini et al. 2009; Bischi et al. 1998), and is out of scope of the current work.

The resulting dynamical system (map) can have one or several fixed points, called Cournot (or Nash) equilibria. It was expected that the mere addition of players would eventually lead to transition of a Cournot equilibrium to a perfect competition equilibrium. However, already in the late 1950s it was shown by Theocharis (1959) [in fact replicating results by Palander (1939) 20 years earlier] that Cournot equilibrium was destabilised when the number of players exceeded a small amount. The same properties were shown to hold by Agiza (1998), Ahmed and Agiza (1998) for a non-linear (isoelastic) demand function and constant marginal costs.

As it was surmised in, e.g., Puu and Ruíz Marín (2006), an assumption that the competitors produce under constant returns seems to be a stumbling block. To resolve this problem, it was suggested to consider eventually decaying returns (see Frisch 1965), modelled by using capacity limits as in Edgeworth (1897). In Puu (2008) it was shown that this can be also achieved by means of so-called constant elasticity of substitution (CES) production functions, with one production factor (capital) being fixed through an act of investment.

As for the market demand function, in most of the studies it is taken in the inverse linear form. Nevertheless, in this work we consider the isoelastic demand function, which seems to be a reasonable choice as it results when the consumers optimise general utility functions of Cobb-Douglas type. Indeed, as known from elementary microeconomics, consumers facing such utility functions spend constant budget shares of their income on each commodity. This means that the quantity of a certain commodity purchased by each consumer is reciprocal to the price of this commodity. Obviously, the sum of all individual demands for this particular commodity is also reciprocal to its price, where the right-hand side constant represents the total value that all consumers together spend on the commodity whose market is studied. Choosing an appropriate price or quantity unit, the constant can always be normalised to unity.

However, the isoelastic demand approach also has its disadvantages. Due to unimodal shape of the derived reaction functions, one is obliged to impose a non-negativity constraint, which results in appearance of a “flat-branch”. As was shown in Tramontana et al. (2010), solutions involving such flat-branches, although being locally unstable (not excluding the totally unstable origin), may become stable in a weak (Milnor 1998) sense. In order to avoid the highly undesirable origin solution, one can stipulate that even if a firm cannot make any profit, it still continues production with some tiny stand-by output, instead of closing down completely. The importance of this numerical value for the resulting dynamics was examined in detail in Tramontana et al. (2010), and its economic meaning was explained as well.

Finally, one has to explicitly devise a dynamical process through which new competitors are added to the market. A way to do this was proposed in Puu (2005), where the firms were supposed to choose new capital stocks, i.e. capacities, at the end of the investment periods when the capital stocks were worn out. As a rule, each new choice of a capital stock is different from the previous, due to changing market conditions. In this way a mixed short/long run dynamical process is set up.

Similar model was already investigated in Panchuk and Puu (2009, 2010), where firms face constant returns in those periods when they renew their capital, but are subject to decreasing returns in the intervening periods when capital is fixed and provides a capacity limit. However, these studies focused on an exogenous process of shifting between constant and decreasing returns. To our knowledge, multi-dimensional models of such kind were poorly studied up to nowadays. In this connection we may mention (Cánovas and Puu 2010), where the authors considered similar approach of introducing the regular investment periods, but using the inverse *linear* demand function and *fixed coefficient technology* production functions.

In the current chapter we continue our studies by suggesting a modified model where the decisions to renew capital stock are endogenous, and depend on how heavily capital equipment has been utilised.

2 Preliminaries

2.1 Demand and Supply

Denoting total market demand (equal to supply in equilibrium) as Q, and price of the good as p, we assume the demand function is of the following *isoelastic* form

$$\sum_{i=1}^{n} q_i = Q = \frac{1}{p}, \tag{1}$$

where n is the total number of competitors in the market, and q_i represents the individual supply of the ith one. Equivalent form of the total demand (supply) may be written as

$$Q = q_i + \sum_{j=1, j\neq i}^{n} q_j = q_i + Q_i \tag{2}$$

where Q_i is called a *residual supply*, which cannot be controlled by the ith firm itself.

The relation (1) results whenever the consumers maximise utility functions of the *Cobb-Douglas* type. Indeed, let a consumer k use a utility function

$$U = \prod_{l=1}^{M} x_{lk}^{\alpha_{lk}}, \qquad \sum_{l=1}^{M} \alpha_{lk} = 1,$$

for choosing between M different commodities. Here $x_{lk} \geq 0$ is the quantity of the l-th commodity and $0 < \alpha_{lk} < 1, l = 1, \ldots, M$ are Cobb-Douglas exponents. Then solving the simple optimisation problem one obtains that the optimal quantity of the

specified good l to be purchased by the consumer k is

$$x_{lk} = \frac{\alpha_{lk} I_k}{p_l},$$

where I_k is the person's income and p_l is the related price. The total quantity demanded by **all** consumers on this chosen good is

$$Q = \frac{\sum_k \alpha_{lk} I_k}{p_l},$$

where the constant $\sum_k \alpha_l I_k$ may be normalised to unity giving directly (1) when dropping the subindex l at p.

It should be mentioned, that there is another popular shape of the demand function which is the linear one $Q = a - bp$. However, it has two disadvantages in comparison with the isoelastic form: (i) one must impose a constraint $p \leq \frac{a}{b}$, in order to prevent demand from becoming negative, and (ii) there is an aggregation issue, namely, the total market demand attains, typically, different form with respect to individual demand functions.[1]

2.2 Production and Cost

Assume the competitors produce using a technology represented by *constant elasticity of substitution* (CES) functions, as written in general form (see, for instance, Arrow et al. 1961)

$$q_i^{-\rho} = A\left(\delta k_i^{-\rho} + (1-\delta)\, l_i^{-\rho}\right), \tag{3}$$

where ρ, A and δ are some positive constants. The quantity units in which we measure output q_i and inputs (capital k_i and labour l_i) are arbitrary, so through a trivial linear change of coordinates we get rid of the constants A and δ and reduce (3) to the symmetric form

$$q_i^{-\rho} = k_i^{-\rho} + l_i^{-\rho}. \tag{4}$$

[1] Indeed, take two consumers with individual demands $x_1 = a_1 - b_1 p$ and $x_2 = a_2 - b_2 p$. In addition, to avoid negative demand, it is required that $x_i = 0$ as soon as $p > \frac{a_i}{b_i}$, $i = 1, 2$. Therefore both individual demands are piecewise linear functions with a single kink point. Consequently, the aggregate (market) demand consists of two sections with different slopes and a third zero-section, hence, it is a piecewise linear function with two kink points. Accordingly, marginal revenue then becomes discontinuous even in a feasible interval for supply, which may lead to existence of several local optima. Obviously, this problem is resolved if all consumers are identical, but this would be rather unlikely.

As for the value of ρ, in most uses of the CES function it is assumed $-1 < \rho < 0$, with the isoquants meeting the axes at tangency, while for $\rho < -1$, the convexity does not make economic sense. In Arrow et al. (1961) the case $\rho > 0$ is also briefly discussed. However, the authors claim that the isoquants then go asymptotically to the axes, which is incorrect. Actually, the asymptotes are located at some distance from the axes, at $k = q$ and $l = q$, in the positive quadrant (see, e.g., Heathfield and Wibe 1987). Taking into account that these asymptotes can be used as capacity limits, in what follows we assume $\rho > 0$. Furthermore, we can specify **any** positive value, as in the topological sense all cases with positive exponents are equivalent. Thus, as long as we are interested in qualitative features rather than numerical exactness, we choose $\rho = 1$ to simplify analytical derivations, which reduces (4) to

$$q_i = \frac{k_i l_i}{k_i + l_i}. \tag{5}$$

Finally, the production costs for the ith firm are

$$C_i = rk_i + wl_i, \tag{6}$$

where $r > 0$ and $w > 0$ denote *capital rent* and *wage rate*, respectively. Assuming that the capital stock k_i is given, solving (5) for labour, and substituting it in (6), one obtains

$$C_i = rk_i + w\frac{k_i q_i}{k_i - q_i}. \tag{7}$$

2.3 *Naïve Expectations and Profits (Long Run)*

To set up a dynamical process one needs to define how the competitors adjust their outputs according to the market situation they observe. Again there can be different approaches, among which we use a rather simple one, often called *naïve expectation* assumption. It means that at every time period each firm has a complete information about what the outputs of all the rivals are. Then it myopically presumes that in the next time period all the other outputs will be the same (which is definitely not true in general). According to this belief, the firm adjusts its output for the next time period so that to comply with its primary task, namely, maximising its profit, which is the difference between the related revenue and costs. Recalling that the ith total revenue is given by

$$R_i = pq_i = \frac{q_i}{Q} = \frac{q_i}{q_i + Q_i}, \tag{8}$$

where Q_i is the residual supply (2), we find the profit of the ith firm as

$$\Pi_i = R_i - C_i = \frac{q_i}{Q_i + q_i} - \left(rk_i + w\frac{k_i q_i}{k_i - q_i} \right). \tag{9}$$

Maximising (9) is a trivial optimisation problem, however, the result depends on whether k_i is assumed to be constant or not. Indeed, let us suppose the capital is **not fixed**, and thus, the firm may modify k_i so that to minimise the costs[2] (7). For this one should take the derivative of (7) with respect to k_i and equate the result to zero, which gives the optimal amount of capital

$$k_i = \left(1 + \sqrt{\frac{w}{r}} \right) q_i. \tag{10}$$

Substituting (10) into (9) one gets

$$\Pi_i^* = \frac{q_i}{Q_i + q_i} - \left(\sqrt{r} + \sqrt{w} \right)^2 q_i = \frac{q_i}{Q_i + q_i} - cq_i, \tag{11}$$

where $\sqrt{c} = \sqrt{r} + \sqrt{w}$. Maximising then (11) for q_i leads to

$$q_i' = \sqrt{\frac{Q_i}{c}} - Q_i, \tag{12}$$

which is the best response of the ith competitor to the expected residual output (here the symbol $'$ stands for the value in the next period). This is what we call a *long run* reaction function.

2.4 Cournot Equilibrium: Local (In)Stability Issue

As for the Cournot equilibrium, note that the firms are distinct only in terms of different capital stocks, and therefore the long run equilibrium implies identical firms. Then

$$q_i = \frac{Q^*}{n}, \qquad Q_i = \frac{n-1}{n} Q^*, \tag{13}$$

[2]Note, that in (9) only the second term C_i, which enters the formula with the minus sign, depends on the capital k_i. Thus, maximising the whole expression means also to minimise the term which is subtracted, namely, the cost function C_i.

where Q^* denotes the total equilibrium supply. Substituting (13) into (12), one gets

$$Q^{*2} = \frac{n-1}{nc} Q^* \quad \text{or} \quad Q^* = \frac{n-1}{nc},$$

and subsequently,

$$q_i = q^* = \frac{n-1}{n^2 c}, \quad k_i = k^* = \frac{n-1}{n^2\sqrt{cr}}, \quad i = 1, \ldots, n. \tag{14}$$

Focusing on local stability of the Cournot equilibrium, one faces the mentioned above Theocharis problem (Theocharis 1959). Though the threshold is $n \geq 4$ in this case (as also noted in Ahmed and Agiza 1998; Agiza 1998). It is a matter of technical computation to construct the Jacobian matrix at the equilibrium point, as well as to derive its eigenvalues which are

$$\lambda^*_{1,\ldots,n-1} = \frac{1}{2}\frac{n-2}{n-1}, \quad \lambda^*_n = -\frac{1}{2}(n-2).$$

The multiple eigenvalue is always less than one in absolute value, and thus never causes any problem, while the last eigenvalue becomes $\lambda^*_n = -1$ for $n = 4$, and Cournot equilibrium is no more attracting for $n \geq 5$.

Intuitively, this problem arises due to the hidden contradiction between modelling the firm's costs as a linear function of its output and the idea of perfectly competitive market. As it is seen from (11), marginal costs are constant, which is equivalent to the production under **constant returns**. It means that the firm is potentially infinitely large in terms of capacity. Meanwhile, one expects that with increasing the number of firms the oligopoly market is transformed to the perfectly competitive market, in which the firms are supposed to be so small, that they cannot affect market price by varying their output. In other words, according to the Amoroso's formula, one has the marginal revenue

$$MR_i = \frac{\partial}{\partial q_i}(pq_i) = p\left(1 + q_i \frac{\partial p}{\partial q_i}\right) \approx p,$$

taking into account that $\frac{\partial p}{\partial q_i}$ is almost zero. Then the ith profit $\Pi_i = R_i - C_i = (p-c)q_i$ is proportional to the output q_i, so theoretically the firm may increase its profit without any limit through increasing its scale of operation. Obviously, if it does so, the price changes, but this is **contradictory** to the idea of the perfect competition, when it is expected that the firms cannot influence the market price whatever amount they produce. The consequent destabilisation of an equilibrium point through adding new firms of infinite size is not very surprising. For studying transition from oligopoly to perfect competition one should rather compare the case with **a few large firms** to the one with **many small firms**. Thus, to derive a consistent model it is necessary to introduce eventually decreasing returns

(capacity limits), excluding the possibility to bring up production unboundedly. One possible approach to avoid this stumbling block was suggested, e.g., in Puu (2005), and used further in Panchuk and Puu (2009, 2010).

2.5 Capacity Limits (Short Run)

To resolve the mentioned above inconsistency of adding to the market new and new infinite sized firms, we introduce *capacity limits* for the short run dynamics. Namely, the capital stock is assumed to be constant for certain number of time periods, then it naturally limits the amount of the produced output, as the function (7) has a vertical asymptote at $q_i = k_i$. Of course, the capital stocks/capacity limits can be different for different firms of the industry. Several sample graphs of short run cost functions together with the long run cost cq_i are plotted in Fig. 1. The points $\hat{q}_1$, $\hat{q}_2$, and $\hat{q}_3$ denote the output values related through (10) to k_1, k_2, and k_3, respectively. Note, that $\hat{q}_i$ are the tangency points of the corresponding short run costs and the long run cost. They may be also called optimal in the following sense: as long as the firm produces less than $\hat{q}_i$, it would be advantageous for it to choose a lower capacity (amount of capital), while if it produces more than $\hat{q}_i$, it would be advantageous to choose a higher capacity (see, e.g., Puu 2008).

To maximise the profit Π_i for q_i, provided that k_i is **fixed**, it is enough to take the derivative of (9) with respect to q_i, and equate it to zero, which implies

$$\frac{Q_i}{(Q_i + q_i)^2} = w\frac{k_i^2}{(k_i - q_i)^2},$$

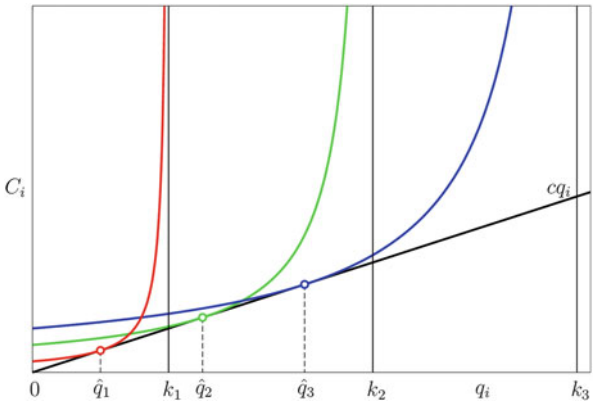


Fig. 1 Sample graphs of short run cost functions with three different capacity limits k_1, k_2, and k_3, together with the long run cost cq_i (*solid slanted line*)

from where we get the reaction function in the *short run*

$$q_i = k_i \frac{\sqrt{\frac{Q_i}{w}} - Q_i}{k_i + \sqrt{\frac{Q_i}{w}}}. \tag{15}$$

Again one may check the local stability of the Cournot equilibrium, whose coordinates are given in (14), by deriving the Jacobian matrix and computing its eigenvalues:

$$\lambda_{1,\dots,n-1} = \frac{1}{2}\frac{n-2}{\left(\sqrt{\frac{r}{w}}n+1\right)(n-1)}, \quad \lambda_n = -\frac{1}{2}\frac{n-2}{\left(\sqrt{\frac{r}{w}}n+1\right)}.$$

Again $|\lambda_{1,\dots,n-1}| < 1$, while $|\lambda_n| < 1$ as soon as

$$w < \frac{4n^2}{(n-4)^2} r \overset{def}{=} \bar{a}(n) r. \tag{16}$$

Detecting that $\lim_{n\to\infty} \bar{a}(n) = 4$, we conclude that asymptotic stability of the Cournot equilibrium in the short run is guaranteed for $w < 4r$. Nonetheless, it may become unstable if $w > 4r$, and the exact threshold depends on the value of n.

2.6 Non-Negativity of Output

Further, we attract reader's attention to the fact that both reaction functions—the short run (15) and the long run (12)—are of the "hump" (unimodal) shape (see Fig. 2). Therefore, to avoid the output becoming negative one has to redefine the reaction functions for $Q_i > \frac{1}{w}$ (short) and $Q_i > \frac{1}{c}$ (long) so that they do not take unfeasible values. The simplest thing is replacing all negative values of the reaction functions by zero. However, we use instead a tiny positive parameter ε

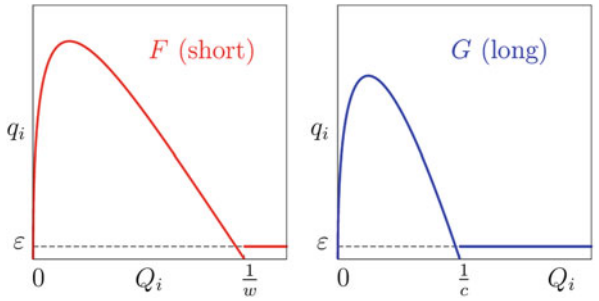


Fig. 2 Short and long run reaction functions

which is called *stand-by output*:

$$\hat{F}(Q,k,w) = \begin{cases} k\dfrac{\sqrt{\frac{Q}{w}} - Q}{\sqrt{\frac{Q}{w}} + k}, & Q < \frac{1}{w} \\ \varepsilon, & Q \geq \frac{1}{w} \end{cases} \tag{17a}$$

$$\hat{G}(Q,c) = \begin{cases} \sqrt{\frac{Q}{c}} - Q, & Q < \frac{1}{c} \\ \varepsilon, & Q \geq \frac{1}{c} \end{cases} \tag{17b}$$

In fact, negativity of the expected output means that the firm has to stop production for the current time period with the possibility to enter the market anew when the situation becomes more auspicious. However, it can happen that closing down completely for a single period and then reopening again involves too much of additional costs (for instance, if the firm uses some equipment which is too expensive to start and stop). Then it is more profitable for the firm to continue production, but generating just a rather small amount of good, which is represented here by the parameter ε. The value of this parameter, though, appeared to have a vital influence on the overall system dynamics (see, e.g., Tramontana et al. 2010).

3 Dynamics of Mixed Short/Long Run Process

3.1 Detecting Periods to Renew the Capital

To finalise the dynamical system formulation, we need to combine now the two reaction functions, setting up a continued iterative process. For that we take into account that the capital wears out with time, and thus, it has to be renewed if production is to be continued (presumably at a quantity different from the previous). Hence, one may construct the mixed short/long run dynamics as follows:

- while the capital of the ith firm is still operable, it produces using the short run reaction function (17a);
- when the capital equipment k_i wears out, the firm buys a new one in the amount calculated according to (10), where the desired output is obtained by using the long run reaction function (17b).

3.1.1 Exogenous Investment Decision

The only thing remained is to define the time periods when the capital is considered to become unfitted. One of the possible approaches (see, e.g., Cánovas and Puu 2010; Panchuk and Puu 2009, 2010) is to make reinvestment decisions exogenous by means of choosing an indicator function $\sigma(i,t)$ such that the ith competitor has to purchase a new capital whenever $\sigma(i,t) = 0$. The function used in the mentioned references has the form

$$\sigma(i,t) = t - mi \mod T, \tag{18}$$

where m is the positive integer, denoting the interval between the initial time periods in which the firms enter the market at the first time, and T is the durability of capital. The relation between m, n, and T defines now how synchronous the competitors are in deciding about renewing their capitals.[3] Below we recall briefly main results concerning the local stability of the Cournot equilibrium in case of exogenous reinvestment decision with the indicator function (18). Namely, we distinguished two particular cases of the capital durability:

1. $T = 2$, "weak" capital, for which the following statements hold (rigorously proved)
   - if m is odd and the number of firms $n \leq 4$, the equilibrium is asymptotically stable for any r, w;
   - if m is odd and $n = 5$, the equilibrium is asymptotically stable for $w < 100r$;
   - if m is odd and $n = 6$, the equilibrium is asymptotically stable for $w < 36r$;
   - if m is odd and $n \geq 7$, or if m is even, the equilibrium is asymptotically unstable for any r, w.
2. $T \geq mn$, "strong" capital, in which case the studies are mainly numerical, and the main observations are
   - if $r = w$, the equilibrium is always asymptotically stable (rigorously proved)
   - in general, the equilibrium is stable for substantial part of the (r, w)-parameter plane, which is bounded by a certain line $w = L(n,T)r$, where $L(n,T)$ is some constant dependent on n and T (numerical evidence).

3.1.2 Endogenous Investment Decision

In what follows we consider another approach for modelling the capital adjustment moments, so that they are endogenous in the dynamic process itself. Namely, for each competitor, besides the two variables q_i and k_i, we introduce another additional

[3] In fact, the main aim of the parameter m was to prevent the situation where all the firms invested at the same period, because the more they synchronise, the greater are the chances that the equilibrium becomes unstable.

real variable T_i representing the remaining time during which the old capital is still usable (in other words, the *current individual durability* of the capital). At each iteration this variable decreases according to a certain law, and when it becomes zero or negative it is supposed that the capital is worn out. In fact, one is free to choose any decaying rule, but we put

$$T_i' = T_i - \kappa^{q_i - q_i^{\text{opt}}}, \tag{19}$$

where the parameter $\kappa \geq 1$, and the value $q_i^{\text{opt}} = \sqrt{\frac{r}{c}} k_i$ is the optimal production value for the current amount of capital k_i. The expression (19) means that the more intensively a firm uses its capital equipment, the quicker it depreciates. As soon as $T_i \leq 0$, the old capital wears out and the new capital has to be purchased, and thus, the long run functions are applied. The parameter κ measures the 'perceptivity' of capital to its overuse or underuse. For instance, if $\kappa = 1$, the capital is not sensitive to whether it is used to optimal extent or not, and whatever amount is produced the capital lifetime always decreases by one. For $\kappa > 1$, if the current production q_i is equal to the optimal supply value q_i^{opt}, then the capital durability T_i decreases by one, if $q_i < q_i^{\text{opt}}$ (capital is underused) then T_i decreases for the value less than one, and if $q_i > q_i^{\text{opt}}$ (the capital is overused) then T_i decreases for the value larger than one. Thus, if the firm tends to overuse its capital (often produces more than the optimal value q_i^{opt}), it has to renew it more often. On the contrary, if the firm underuses its capital, the capital lives for longer time than it is expected. And the larger is κ, the stronger the capital reacts to the fact of its overuse or underuse.

In such a way, the final map denoted Φ acquires the form:

$$\begin{aligned}
q_i' &= \begin{cases} \hat{F}(Q_i, k_i, w), & T_i > 0, \\ \hat{G}(Q_i, c), & T_i \leq 0, \end{cases} \\
k_i' &= \begin{cases} k_i, & T_i > 0, \\ \sqrt{\frac{c}{r}} \hat{G}(Q_i, c), & T_i \leq 0, \end{cases} \qquad i = 1, \ldots, n, \\
T_i' &= \begin{cases} T_i - \kappa^{q_i - \sqrt{\frac{r}{c}} k_i}, & T_i > 0, \\ T_0, & T_i \leq 0, \end{cases}
\end{aligned} \tag{20}$$

where T_0 denotes the *global durability* of the capital, and the functions $\hat{F}$, $\hat{G}$ are defined in (17). To sum up, the dynamical process may be described as follows. In those time periods when the current individual durability of capital T_i is positive (the old capital is still usable), the related capital value k_i is taken as fixed, and thus, the ith competitor is producing under capacity limit (defined by the value k_i). Then the reaction function is chosen in the short run form (15). As soon as T_i becomes zero or negative, the old capital is considered to be worn out (not usable any more),

hence, the ith competitor has to reinvest—to purchase a new capital stock. In this case, the choice for the output to be produced in the next time period is calculated according to the long run reaction function (12), while the optimal amount of new capital is adjusted according to (10), and current durability T_i of this new capital is set to the predefined value T_0.

It should be mentioned that in Cournot *duopoly* models (subjected to naïve expectations) the asymptotic dynamics is often studied by means of a special composite one-dimensional map, which corresponds in some sense to the second iterate of the two-dimensional system function (see, e.g., Bischi et al. 2000; Tramontana et al. 2010). However, for the number of firms $n \geq 3$, this trick does not work anymore, as there is no explicit separation of variables $q_i, i = 1, \ldots, n$. In addition, the map Φ (20) is $3n$-dimensional, as for each competitor we introduce two supplementary variables, k_i and T_i, which also change depending on the produced amount of good q_i.

In high-dimensional systems it is usually not an easy task to examine how changes in initial conditions may influence asymptotic dynamics of trajectories, especially when a system shows sensitivity to initial conditions. Therefore, for definiteness, in our numerical simulation we use initial conditions which are reasonable from the economic viewpoint. Namely, if not stated otherwise, we assume that each firm when entering the market produces a tiny amount of good ε, and for that chooses an optimal small value of capital, namely,

$$q_i^0 = \varepsilon, \quad k_i^0 = \sqrt{\frac{c}{r}}\varepsilon, \tag{21}$$

while $(T_1^0, \ldots, T_n^0)$ may be arbitrary. And then the starting process is set up as follows: during the first T_i^0 iterates the ith firm produces nothing (is inactive), and then at the step $T_i^0 + 1$ enters the market with small capital k_i^0 producing tiny amount of good q_i^0. The initial values T_i^0, $i = 1, \ldots, n$, are referred to as *initial inactivity times*.

3.2 Regular Short/Long Run Switching: $\kappa = 1$

Let us consider first the case $\kappa = 1$. Then at each iteration the value of T_i, $i = 1, \ldots, n$, decreases by one if being positive, and is set to T_0 otherwise. Although modelled in another way, this case has much in common with the one mentioned in Sec. 3.1.1. For instance, if initial inactivity times $T_i^0 = im$, $i = 1, \ldots, n$, with some positive integer m, the sequel of switching between short and long run moments is exactly the same as in the case with exogenous reinvestment moment definition (the only difference, which one has to keep in mind, is that the global durabilities for exogenous and endogenous cases are related as $T_0 = T - 1$). The stability regions of the Cournot equilibrium with the endogenous choice of capital renewal moments for different T_0 and n are plotted in Fig. 3a–d, which is perfectly compliant with

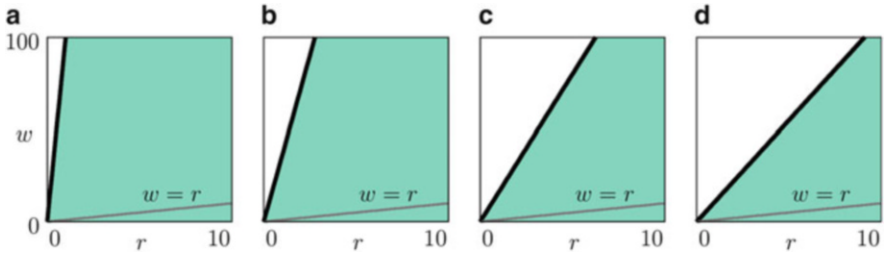


Fig. 3 Stability regions for the Cournot equilibrium in case of endogenously defined reinvestment moments. The parameters are (**a**) $T_0 = 1$, $n = 5$; (**b**) $T_0 = 1$, $n = 6$; (**c**) $T_0 = 15$, $n = 8$; (**d**) $T_0 = 25$, $n = 10$. *Black solid lines* indicate (**a**) $w = 100r$; (**b**) $w = 36r$; (**c**) $w = 16r$; (**d**) $w = 11r$. *Grey line* is the main diagonal $w = r$

the results obtained in the former (exogenous) case. (For obtaining these diagrams the initial conditions were chosen in the equilibrium neighbourhood). Furthermore, concerning the condition (16), it can be found out that $\bar{a}(5) = 100$, $\bar{a}(6) = 36$, $\bar{a}(8) = 16$, and $\bar{a}(10) = 11.111\ldots$, and the boundaries of the regions shown in Fig. 3a–d are close to the lines $w = 100r$, $w = 36r$, $w = 16r$, $w = 11r$, respectively.

This numerical study may be interpreted economically as follows. If the number of competitors increases, the region for the local stability of the Cournot equilibrium shrinks, although it is still present even if the previously declared instability threshold of five (or four in case of linear demand) suppliers is overpassed. However, the greater is the number of firms in the market, the larger should be the global durability of the capital T_0, so that to guarantee that in each time moment not more than a single firm renews its capital. Having in mind that the firms renewing their capitals use the long run function, and hence, produce under constant returns (are potentially infinitely large), this conclusion is not much surprising.

In addition, one may also raise a hypothesis that the overall dynamics depends only on the ratio between r and w. Indeed, introducing the change of variables

$$\tilde{q}_i = wq_i, \quad \tilde{k}_i = wk_i, \quad i = 1, \ldots, n, \tag{22}$$

an appropriate rescaling of the parameters of (20) may be performed, which leads to reduction of the number of parameters by one with keeping the unique parameter $\sqrt{\frac{w}{r}}$. However, for more suitable interpretation of the system behaviour in the applied context we prefer to avoid the rescaling (22). Instead (having in mind that the qualitative dynamics of the map (20) depends only on the ratio $\frac{w}{r}$) we fix the wage rate value $w = 1$ and also denote $\hat{F}(Q_i, k_i, 1) \stackrel{def}{=} F(Q_i, k_i)$, $\hat{G}(Q_i, \sqrt{r} + 1) \stackrel{def}{=} G_r(Q_i)$.

One of the peculiarities of the map (20) is that any point of the form $(q^*, \ldots, q^*, k^*, \ldots, k^*, T_1, \ldots, T_n)$, with q^*, k^* given in (14), corresponds to the Cournot equilibrium. However, even if $\mathbf{q} = (q_1, \ldots, q_n)$ and $\mathbf{k} = (k_1, \ldots, k_n)$ are fixed, the last n variables $\mathbf{T} = (T_1, \ldots, T_n)$ still continue to change at each iteration. Thus, for the map Φ for any $T_0 > 0$

1. there are $(T_0 + 1)^n$ points representing the Cournot equilibrium;
2. all these points are not fixed points, but *periodic* ones with period $T_0 + 1$.

3.2.1 Reinvestment Synchronisation Manifolds

In general, asymptotic dynamics of the map Φ trajectories depends essentially not only on the parameter T_0, but also on the initial inactivity times $\mathbf{T}^0 = (T_1^0, \ldots, T_n^0)$. Indeed, depending on the initial distribution of T_i's and the value of T_0, the competitors may merge into groups which fall back into long run simultaneously. For instance, if $n = 6$, $T_0 = 2$, and $\mathbf{T}_1^0 = (2, 4, 6, 8, 10, 12)$, after several iterations the vector of current capital durabilities $\mathbf{T}$ will jump cyclically between three different vectors $\mathbf{T}_1 = (1, 0, 2, 1, 0, 2)$, $\mathbf{T}_2 = (0, 2, 1, 0, 2, 1)$, and $\mathbf{T}_3 = (2, 1, 0, 2, 1, 0)$. Note, that due to this (i) the period of any cycle of Φ must be *multiple of three*, (ii) after a finite number of iterations any trajectory will be trapped by the manifold $M_1 = \{(\mathbf{q}, \mathbf{k}, \mathbf{T}) \ : \ T_1 = T_4, T_2 = T_5, T_3 = T_6\}$. On the other hand, if $\mathbf{T}_2^0 = (2, 4, 5, 7, 8, 10)$, after several iterations the vector $\mathbf{T}$ will jump cyclically between the vectors $\tilde{\mathbf{T}}_1 = (0, 2, 0, 2, 0, 2)$, $\tilde{\mathbf{T}}_2 = (2, 1, 2, 1, 2, 1)$, and $\tilde{\mathbf{T}}_3 = (1, 0, 1, 0, 1, 0)$. Again, any cycle of Φ is $3 \cdot s$-periodic ($s = 1, 2, \ldots$) and any trajectory asymptotically converges to another manifold $M_2 = \{(\mathbf{q}, \mathbf{k}, \mathbf{T}) \ : \ T_1 = T_3 = T_5, T_2 = T_4 = T_6\}$.

Thus, in case of regular short/long run switching $\kappa = 1$, with fixed n and T_0, the vector of initial inactivity times defines the manifold to which trajectories of the system (20) converge with time. We refer to such kind of manifolds as *reinvestment synchronisation manifolds*.

It is clear that the map Φ is symmetric with respect to the arbitrary renumbering of the state variables, namely,

$$\Phi(q_1, \ldots, q_n, k_1, \ldots, k_n, T_1, \ldots, T_n) \equiv \Phi(q_{i_1}, \ldots, q_{i_n}, k_{i_1}, \ldots, k_{i_n}, T_{i_1}, \ldots, T_{i_n}),$$

where $\{i_1, \ldots, i_n\}$ is some permutation of the set $\{1, \ldots, n\}$. Therefore, to study qualitative dynamics (studying global basins of attraction is, surely, a different story, which is rather cumbersome for a $3n$-dimensional system) it is enough to consider reinvestment synchronisation manifolds of the form

$$T_1 = \ldots = T_{l_1}, \ T_{l_1+1} = \ldots = T_{l_2}, \ \ldots, \ T_{l_m+1} = \ldots = T_n, \tag{23}$$

where $m \leq n$ is a certain positive integer.

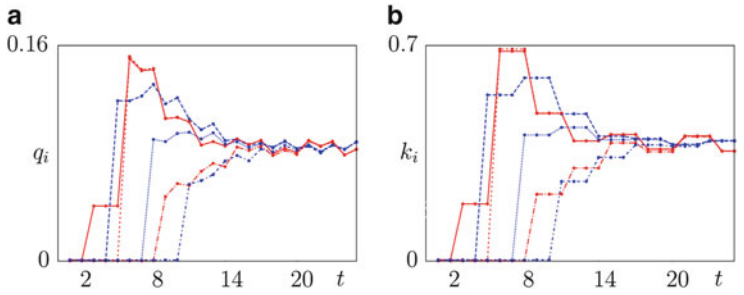


Fig. 4 Time series of (**a**) q_i and (**b**) k_i, $i = 1, \ldots, 6$, which synchronise into two equal-sized clusters. Initial inactivity times are $\mathbf{T}^0 = (2, 4, 5, 7, 8, 10)$, and $r = 0.08, n = 6, T_0 = 2$

Note, that conditions (23) imply also similar equalities (to be satisfied asymptotically) for q_i and k_i, $i = 1, \ldots, n$, if initial values q_i^0, k_i^0 are taken in the form (21). This is illustrated in Fig. 4 (where different line styles correspond to different competitors), and can be explained as follows. First, only one firm is active, and adjusts its output and capital to take on optimal values with respect to tiny production ε of the others. Then the second firm enters the market ($t = 4$) and produces some larger amount of good. The first firm then reacts with increasing the amount of its capital at $t = 5$. However, at the same time ($t = 5$) the third firm appears and chooses amount of capital also being close to maximal value (this directly follows from the form of the long run reaction function G_r, see Fig. 2). From now on, the first and the third firms react in synchrony producing the same output, because the time moments at which they decide about changing their capital coincide. In a similar way all firms split into two synchronised groups after a small number of iterations. Thus, provided that conditions (23) are fulfilled, the following equalities are also held asymptotically

$$\begin{aligned} &q_1 = \ldots = q_{l_1},\ q_{l_1+1} = \ldots = q_{l_2},\ \ldots\ q_{l_m+1} = \ldots = q_n, \\ &k_1 = \ldots = k_{l_1},\ k_{l_1+1} = \ldots = k_{l_2},\ \ldots\ k_{l_m+1} = \ldots = k_n. \end{aligned} \tag{24}$$

3.2.2 Full Synchronisation

To shed light on the phenomena observed when studying dynamics of the map (20), we consider first the simplest case, namely, when all the firms reinvest (asymptotically) in the same time moment. It means $T_1 = \ldots = T_n$ that entails also $q_1 = \ldots = q_n$, $k_1 = \ldots = k_n$, which defines the domain in the state space denoted as M_C. In this case, during T_0 successive iterates all firms choose the short run function F, and at $T_0 + 1$ iterate they all use the long run function G_r. The dynamics of the map Φ (20) on M_C can be reduced to a three-dimensional map

$\Psi(q,k,T)$ as follows

$$q' = \begin{cases} F((n-1)q,k), & T > 0, \\ G_r((n-1)q), & T \le 0, \end{cases}$$

$$k' = \begin{cases} k, & T > 0, \\ \left(1+\dfrac{1}{\sqrt{r}}\right) G_r((n-1)q), & T \le 0, \end{cases} \tag{25}$$

$$T' = \begin{cases} T - \kappa^{q-q^{\text{opt}}}, & T > 0, \\ T_0, & T \le 0. \end{cases}$$

The characteristic feature of the map given in (25) is presence of a "flat part" defined by the parameter ε (here it is the plane $q = \varepsilon$). As it was already mentioned, influence of the numerical value ε on the overall asymptotic dynamics of such kind systems was explored in Tramontana et al. (2010). Obvious parallels may be drawn between the phenomena described in the mentioned reference and the asymptotic behaviour of the map Ψ. However, in our case the situation is complicated by the fact that at every $T_0 + 1$ iterate

- we have a "perturbation" in the sense that we jump from the function F to the function G_r, and
- the form of the function F changes as it depends also on the modified value of the capital k.

If the number of firms $n \le 5$ then the domain

$$\Pi = \left[0, \frac{1}{(1+\sqrt{r})^2}\right] \times \left[0, \frac{1}{4\sqrt{r}(1+\sqrt{r})}\right] \times [0, T_0] \tag{26}$$

is the trapping area, and thus, any trajectory of the map Ψ after a finite number of iterates enters Π and then stays there forever. This is due to the fact that for $n \le 5$

$$\begin{aligned} \max_{q,k} F((n-1)q,k) = \max_{q} G_r((n-1)q) &= \frac{1}{4(1+\sqrt{r})^2} \\ &\le \frac{1}{(n-1)(1+\sqrt{r})^2} < \frac{1}{n-1}. \end{aligned} \tag{27}$$

If $n \ge 6$, the condition (27) does not hold, therefore some trajectories may leave Π after a finite number of iterates, and be attracted to a cycle having a point on the flat part

$$\Pi_f = \left\{ (q,k,T) \,:\, q = \varepsilon,\ 0 \le k \le \frac{1}{4\sqrt{r}(1+\sqrt{r})},\ 0 \le T \le T_0 \right\}.$$

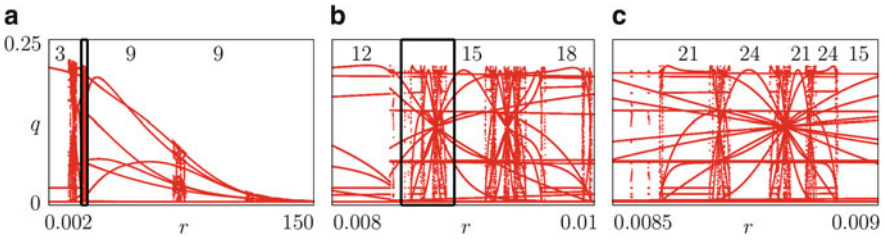


Fig. 5 One-dimensional bifurcation diagram for the system (25) with $T_0 = 2$, $\varepsilon = 0.0001$. The graph (**b**) shows a zoomed *rectangle outlined black* in (**a**), and (**c**) shows a zoomed rectangle outlined black in (**b**)

In Fig. 5a a typical 1D bifurcation diagram for the map (25) is plotted with $n = 6$ and $T_0 = 2$, while Fig. 5b is a zoom of the region outlined black in Fig. 5a, and Fig. 5c is a further zoom of the region marked by black line in Fig. 5b. The numbers at the top of the graph denote periods of the underlying cycles. As one can see, the bifurcation structure is self-similar and has infinitely many "spider-like" nodes, more and more of which show up when zooming. These patterns consist of the cycles which appear and disappear through a border collision bifurcation and all have a point belonging to the flat part Π_f (this phenomenon is similar to that described in Tramontana et al. 2010, 2011).

Furthermore, every period is a multiple of $T_0+1 = 3$ (as mentioned above), however the principle, according to which period changes with the varying parameter, is not so obvious. From Fig. 5, it is seen that on one side of each "spider-node" the cycle periods are odd, and on the other side they are even, more precisely, $3 \cdot 2s$ and $3 \cdot (2s + 1), s = 1, 2, \ldots$. However, it is also noticeable that between any two nodes there is another one (in fact, infinitely many ones), and therefore the sequences of odd and even periods are highly intermingled. As a result, it is difficult to predict the asymptotic dynamics of the system, as small changes in parameter values may cause abrupt modification of the map trajectories.

Additionally, as each solution inevitably has a point $q = \varepsilon$, one may surmise that such dynamics is not fine from the economic viewpoint. Really, turning back to the full-space model (20), $q = \varepsilon$ implies that $q_1 = \ldots = q_n = \varepsilon$, corresponding to $Q = n\varepsilon$. It means that periodically firms flood the market by the good produced, and right after that they all have to depress their production to the stand-by output. This may signify the bad adjustment strategy chosen, as well as it also means that the market has substantial latent instabilities. Again it is not surprising, recalling that being in the long run the firm is potentially infinitely large (producing under constant returns). In case when all suppliers react in synchrony, they all renew their capitals at the same time, and in the related time period all the firms in the market are of infinite size (which is similar to the issue uncovered for the Theocharis problem). Thus, from practical viewpoint it is better to avoid situation when all firms synchronise.

3.2.3 Forming Clusters: Dependence on Initial Inactivity Times

In comparison to the full synchronisation one may also consider the case when firms form smaller groups, which then adjust their capitals in the same time periods. For that one needs to choose the appropriate initial inactivity times $\mathbf{T}^0$, e.g., as in examples mentioned above with $\mathbf{T}_1^0 = (2, 4, 6, 8, 10, 12)$ or $\mathbf{T}_2^0 = (2, 4, 5, 7, 8, 10)$. In the first case, after a small number of iterates we observe *three groups of two* synchronised firms, while in the second one, *two groups of three* synchronised firms are formed. Hence, an initial condition may be taken already on the appropriate synchronisation manifold, like $\tilde{\mathbf{T}}_1^0 = (0, 0, 1, 1, 2, 2)$ and $\tilde{\mathbf{T}}_2^0 = (0, 0, 0, 2, 2, 2)$.

It is not surprising that these two cases lead to considerably different asymptotic dynamics, as can be seen from the one-dimensional bifurcation diagrams presented in Fig. 6, where the total market supply Q and the synchronisation rate (formed cluster sizes) are plotted versus the varied parameter r. In Fig. 6c, d different lines indicate number of firms belonging to each synchronised group (cluster). For some region of larger r values all firms form a single cluster with six elements, and it corresponds to the Cournot equilibrium. For the rest of parameter values in the first case one observes three groups of two firms (Fig. 6c), and in the second case there are two groups of three firms (Fig. 6d).

It is noticeable that in the less synchronised market (three clusters, Fig. 6a) the unwanted situation of $Q = n\varepsilon$ arises for narrower range of r, while for the stronger synchronisation (two clusters, Fig. 6b) such kind of solutions still appear

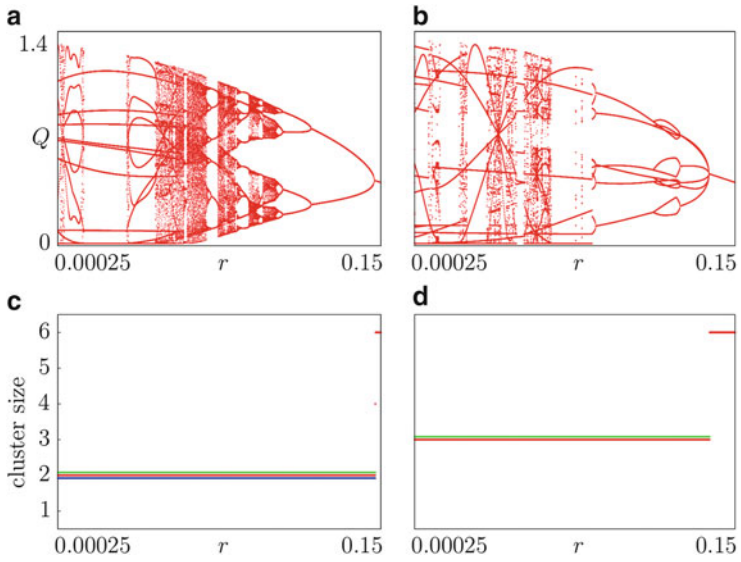


Fig. 6 The total market supply Q (**a**, **b**), and cluster sizes (**c**, **d**) vs. r with initial distribution of capital durabilities (**a**, **c**) $\tilde{\mathbf{T}}_1^0 = (0, 0, 1, 1, 2, 2)$; (**b**, **d**) $\tilde{\mathbf{T}}_2^0 = (0, 0, 0, 2, 2, 2)$. The other parameters are $\varepsilon = 0.0001$, $T_0 = 2$

for larger values of r (compare also with Fig. 5, where the solutions having the point $Q = n\varepsilon$ exist for the whole range of r values). In other words, the behaviour of system (20) depends on the tendency of the competitors to synchronise in changing their capital equipment (reinvest). And intuitively it is clear. Indeed, let several firms with indices $i_1, i_2, \ldots, i_m$ synchronise, and suppose that at some time moment the cumulative amount of good produced is too large, that is $Q_{i_1} = \ldots = Q_{i_m} > 1$ (or $Q_{i_1} > \frac{1}{(1+\sqrt{r})^2}$ if in the long run). As a result, at the next iterate the optimal strategy for these m competitors is to produce a tiny amount of good ε. On the contrary, the remaining $n - m$ firms tend then to produce more, which after a couple of iterates may lead their residual supply to exceed critical value as well. Thus, asymptotically the map Φ dynamics is sensitive to the fact whether the firms have to depress their production simultaneously or not.

The typical two-dimensional bifurcation diagrams in the (r, ε) parameter plane, for $n = 6, T_0 = 2$ and the two initial vectors $\tilde{\mathbf{T}}_1^0 = (0, 0, 1, 1, 2, 2)$ and $\tilde{\mathbf{T}}_2^0 = (0, 0, 0, 2, 2, 2)$, are plotted in Fig. 7. Note, that the region related to the Cournot equilibrium is denoted by 1, although the periodicity of this solution is 3, because the coordinates T_i always change cyclically. As one can see, for larger r-values the bifurcation scenarios in both cases are similar and do not depend on ε, meaning that these solutions do not contain any points on the flat part with $Q = n\varepsilon$ (cf. Fig. 6c, d). On the contrary, for smaller r-values the dynamics in Fig. 7a, b are totally different.

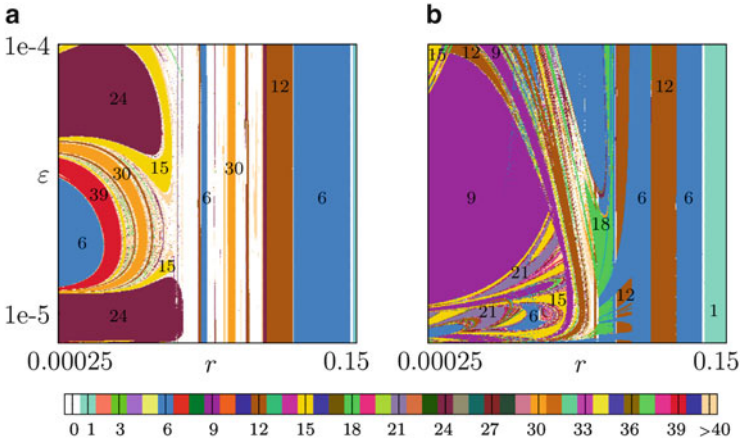


Fig. 7 Typical 2D bifurcation diagrams in the (r, ε) parameter plane. (**a**) $\tilde{\mathbf{T}}_1^0 = (0, 0, 1, 1, 2, 2)$; (**b**) $\tilde{\mathbf{T}}_2^0 = (0, 0, 0, 2, 2, 2)$. $n = 6, T_0 = 2$

3.3 Less Regular Short/Long Run Switching: $\kappa > 1$

Let us now examine how the parameter κ influences the asymptotic dynamics of the map Φ.

And first we focus on the case when the trajectories converge to the full synchronisation manifold M_C. In Fig. 8a, b two-dimensional bifurcation diagrams for the map Φ (25) in the (r, ε) parameter plane are shown for $\kappa = 1$ and $\kappa = 1.1$, respectively. In the first plot the observed dynamics corresponds to the solutions having a point on the flat part $Q = n\varepsilon$, and is totally defined by the trajectory of the point $(\varepsilon, \ldots, \varepsilon, \sqrt{\frac{c}{r}}\varepsilon, \ldots, \sqrt{\frac{c}{r}}\varepsilon, 1, \ldots, 1)$ giving rise to different periodic solutions (cf. Fig. 5). The second plot looks totally different, in particular, for larger values of r the Cournot equilibrium is reached. Note, that the indicated periods are calculated without including the last variables T_i, $i = 1, \ldots, n$. For instance, in the region related to the number 10 the period of the map Φ solution is in fact 30, but the outputs $(q_1, \ldots, q_n)$ jump cyclically between 10 vectors.

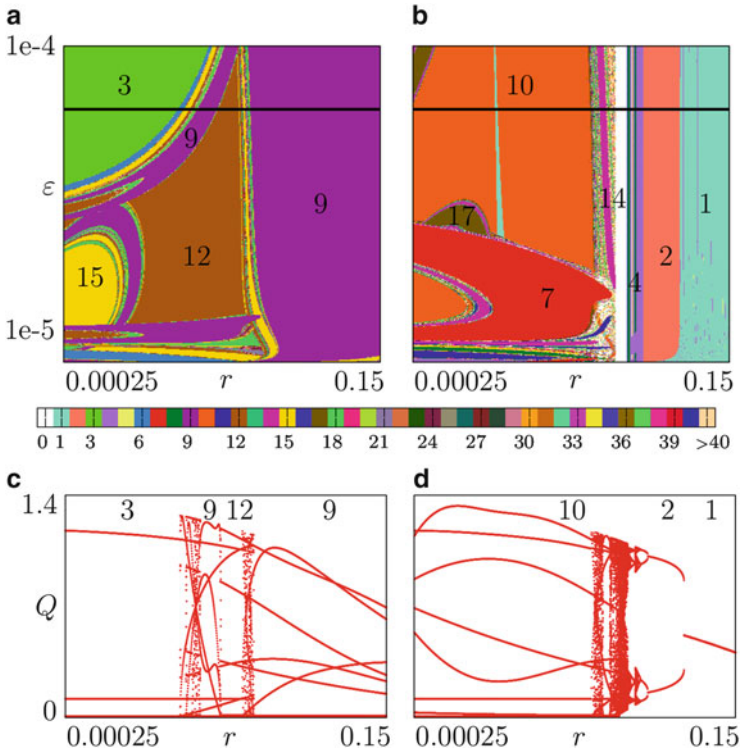


Fig. 8 (**a**, **b**) Two-dimensional bifurcation diagrams for Ψ with (**a**) $\kappa = 1$, (**b**) $\kappa = 1.1$. (**c**, **d**) One-dimensional bifurcation diagrams for Ψ with $\varepsilon = 0.00008$ and (**c**) $\kappa = 1$, (**d**) $\kappa = 1.1$. ($T_0 = 2, n = 6$)

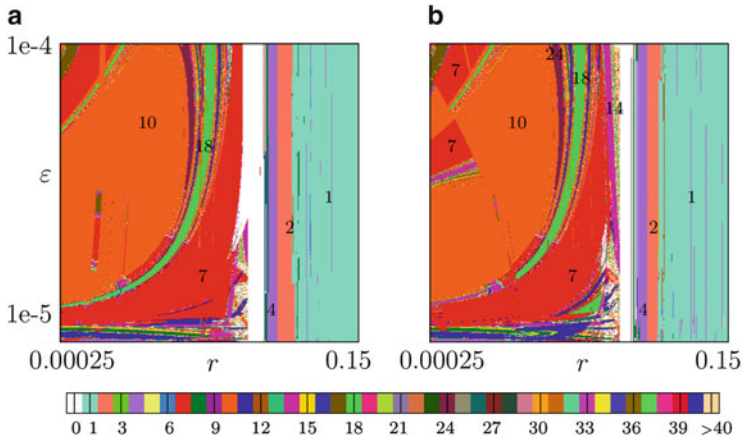


Fig. 9 Two-dimensional bifurcation diagrams for the map Φ with (**a**) $\mathbf{T}^0 = (2, 4, 6, 8, 10, 12)$ and (**b**) $\mathbf{T}^0 = (2, 4, 5, 7, 8, 10)$. $T_0 = 2$, $\varepsilon = 0.0001$, $\kappa = 1.1$

In Fig. 8c, d one-dimensional bifurcation diagrams for $\varepsilon = 0.00008$ are displayed, corresponding to the upper graphs with $\kappa = 1$ and $\kappa = 1.1$. The numbers at the top of the graph again denote periodicities of the underlying cycles. As one can see, for less regular switching between short and long run ($\kappa > 1$) the "bad" solution having point on the flat part Π_f appears only for smaller r values. (In contrast to the case of regular switching, $\kappa = 1$, when it is reached for the whole range of r).

As for the case when the firms split into several groups which reinvest in synchrony, the result is different. In Fig. 9a, b we show two-dimensional bifurcation diagrams for $\kappa = 1.1$ with initial inactivity vectors $\mathbf{T}_1^0 = (2, 4, 6, 8, 10, 12)$ and $\mathbf{T}_2^0 = (2, 4, 5, 7, 8, 10)$, respectively. As it can be seen, the two plots have much in common. The reason why it happens gets clear when looking at the related cluster sizes shown in Fig. 10c, d. Namely, both graphs look similar, and one can see that the firms either form a single cluster (full synchronisation), or they split into two groups.

In particular, for $\mathbf{T}_1^0$ increasing κ has surged also the rate of synchronisation in the sense that instead of three clusters there are now two (or even a single one). Thus, as visible from Fig. 10a where the corresponding one-dimensional bifurcation diagram for Q vs. r is displayed, the "bad" solution (having point on the flat part $\{q_1 = \ldots = q_n = \varepsilon\}$) appears for larger parameter range in comparison to Fig. 6a. On the other hand, the interval where trajectories are attracted to Cournot equilibrium is also enlarged.

As for $\mathbf{T}_2^0$, the range of r, for which the "bad" solution with $Q = n\varepsilon$ occurs, remains almost the same, although the rate of synchronisation increases as well (see Fig. 10d). And the interval where Cournot equilibrium shows up is enlarged. Obviously, the overall dynamics changes qualitatively showing regular solutions of periods different from the case $\kappa = 1$ (cf. Figs. 10b and 6b).

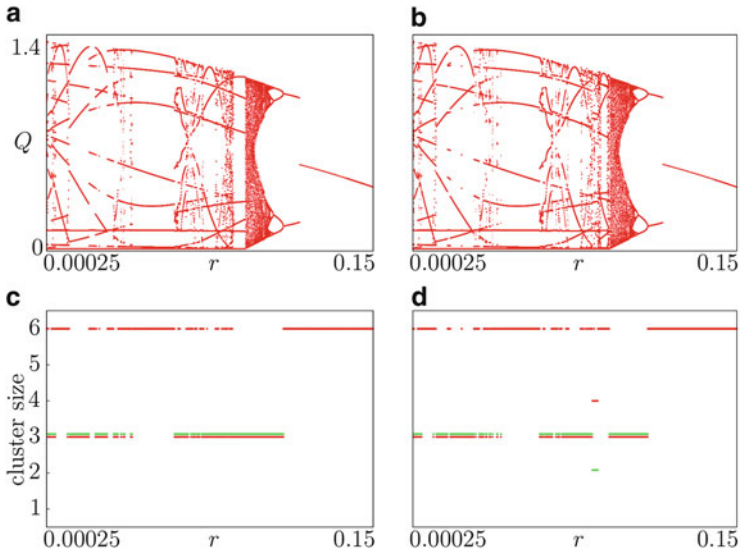


Fig. 10 One-dimensional bifurcation diagrams for the map Φ (*upper panels*) and sizes of synchronised groups of firms (*lower panels*). The initial capital durabilities are (**a**, **c**) $\mathbf{T}^0 = (2, 4, 6, 8, 10, 12)$; (**b**, **d**) $\mathbf{T}^0 = (2, 4, 5, 7, 8, 10)$. The other parameters are $T_0 = 2$, $\varepsilon = 0.0001$, $\kappa = 1.1$

In other words, with putting $\kappa > 1$ the market situation may get worse—acquiring more hidden instabilities, and thus, inducing larger parameter range for which "bad" solution with $Q = n\varepsilon$ is attracting. On the other hand, for certain parameter choice the market situation may also get better when κ changes from unity to the larger value, namely, some hidden instabilities are suppressed and attracting "bad" solution appears for narrower parameter range. It seems that whether $\kappa > 1$ is advantageous or not (in the sense mentioned above) depends mostly on the value of r and initial inactivity times vector $\mathbf{T}^0$. This again reverts us to the hypothesis that the more firms reinvest in synchrony, and the more often the long run function is used, the stronger is the latent imbalance of the market (so that it is more severely affected by potential infiniteness of the firms producing under constant returns).

4 Concluding Remarks

This chapter presented an overview of certain results concerning investigation of an oligopoly market using CES production function in combination with the isoelastic demand function and introducing capacity limits. The concept of the model described emerged from the attempt to resolve the problem of destabilisation of the Cournot equilibrium when the number of suppliers exceeded a very small threshold (ascribed to Theocharis) which upset the economists so much. This issue makes

it difficult to believe that when more competitors are added, the market undergoes a series of transformations from monopoly over duopoly and oligopoly to perfect competition (which is the ideal socially optimal case favouring the consumers). Even if the Cournot equilibrium price approaches the perfect competition price, this equilibrium lacks for interest if it is unstable.

Some time ago it was suggested by T. Puu that the Theocharis problem was misconceived, in the sense that considering increasing competition under constant returns means just adding to the market more and more firms of infinite production potential. However, the main idea of perfect competition supposes the firms are so small ("atomistic") that they cannot affect the market price irrespective of how much they produce in their possible range of the output variation. Now, this case is contradictory with the case of constant returns to scale: atomistic firms can never be at the same time infinitely large in production potential. And then there came out an idea to emulate the firms with decreasing returns, or even capacity limits. It appeared to be possible to derive the appropriate model from the CES family of functions with subsequent solving for the short run reaction functions. Then a very stylised mixed short/long run dynamical process was set up, where each individual firm had to choose a new capital stock at the end of each investment period (long run), hence acting under constant returns, while in the intervening periods (short run) its capital was fixed, and thus, provided a capacity limit.

Such a dynamical process was studied within a certain aspect in the present chapter. The definition of the moment at which it is time to renew the capital was considered in two possible variants: it may be either exogenous, or endogenous in the model itself. In the latter case there was also examined a possibility when the lifetime of the capital equipment depends on how heavily it has been utilised.

Having in mind that a firm producing under constant returns is potentially infinitely large, it is not surprising that the more such firms are present in the market, the more noticeable local instabilities are. For instance, it was numerically demonstrated that if the number of competitors increases, the region for the local stability of the Cournot equilibrium shrinks. However, this region is still present even if the previously declared instability threshold of five (or four in case of linear demand) suppliers is overpassed, although one has to make sure that the global durability of the capital T_0 is sufficiently large.

Another observation is that the asymptotic dynamics on the market in general depends on the value of the global capital durability T_0. In particular, when the equilibrium point is unstable, one may observe instead certain attracting periodic solutions whose period must be a multiple of $T_0 + 1$. Furthermore, the behaviour of the system is sensitive to the initial choice of individual inactivity times $\mathbf{T}^0 = (T_1^0, \ldots, T_n^0)$, that defines in fact, the periods at which the firms activate their production. Different combinations of $\mathbf{T}^0$ and values of T_0 may lead to various types of synchronisation between the competitors, namely, they merge into groups which renew their capitals simultaneously. And this fact influences much the overall dynamics of the considered map. Which is again not much surprising from the economics viewpoint, because the more suppliers synchronise, the more of them renew their capitals at the same time period. It means that in this particular period

certain firms in the market produce under constant returns and are potentially of infinite size. The more firms comply to this criteria simultaneously, the stronger are the latent instabilities in the market. The extreme case of such instabilities becoming apparent is existence of the attracting solution having a point on the flat branch, namely, when $Q = n\varepsilon$. This is what we call a "bad" solution, which we would prefer to avoid, as it disrupts the smoothness of economic interpretation.

Finally, we also investigated how the system behaviour changes with varying the parameter κ which controls the rate of sensitivity of the capital to its overuse or underuse. It is shown that the influence of this parameter is ambiguous: for some parameter sets, putting $\kappa > 1$ may improve the situation on the market, in the sense that this may lead to stabilisation of the Cournot equilibrium, or at least force the "bad" solution $Q = n\varepsilon$ to appear for the smaller parameter range (in comparison to the case of $\kappa = 1$). On the other hand, for certain parameter configurations, $\kappa > 1$ may spoil things forcing the firms to synchronise more, which leads to more frequent occurrence of the "bad" kind of dynamics, when periodically all firms has to stop their production and simultaneously vacate the market, starting all over again in the next time period. All these observations allow us to make a hypothesis that the more firms reinvest in synchrony, and the more often the long run function is used, the stronger are the latent instabilities in the market (so that it is more severely affected by potential infiniteness of the firms using the long run reaction function).

As one of the possible further developments of the model presented one may consider modified reaction functions. In fact, the best replies obtained under assumption of naïve expectations are just conjectures. No competitor can be sure about the future, so it is prudent to move *only part* of the way from the previous choice towards the calculated best reply. For that one may introduce a certain adaptation coefficient whose aim is to regulate how strongly the supplier relies on the solution obtained from the profit maximisation problem. Such an adaptive form is considered by the current author together with the two other colleagues Prof. Jose Cánovas and Prof. Tönu Puu in the frame of research which is in progress.

References

Agiza, H. N. (1998). Explicit stability zones for Cournot games with 3 and 4 competitors. *Chaos Solitons Fractals*, *9*, 1955–1966.

Agliari, A. (2006). Homoclinic connections and subcritical Neimark bifurcations in a duopoly model with adaptively adjusted productions. *Chaos Solitons Fractals, 29*, 739–755.

Agliari, A., Gardini, L., & Puu, T. (2002). Global bifurcations of basins in a triopoly game. *International Journal Bifurcation Chaos, 12*, 2175–2207.

Ahmed, E., & Agiza, H. N. (1998). Dynamics of a Cournot game with *n* competitors. *Chaos Solitons Fractals, 9*, 1513–1517.

Ahmed, E., Agiza, H. N., & Hassan, S. Z. (2000). On modifications of Puu's dynamical duopoly. *Chaos Solitons Fractals, 11*(7), 1025–1028.

Angelini, N., Dieci, R., & Nardini, F. (2009). Bifurcation analysis of a dynamic duopoly model with heterogeneous costs and behavioural rules. *Mathematics and Computers in Simulation, 79*, 3179–3196.

Arrow, K. J., Cheney B., Minhas, B., & Solow, R. M. (1961). Capital-labor substitution and economic efficiency. *Review of Economics and Statistics, 43*, 225–250.

Bischi, G.-I., Stefanini, L., & Gardini, L. (1998). Synchronization, intermittency and critical curves in a duopoly game. *Mathematics and Computers in Simulation, 44*, 559–585.

Bischi, G. -I., Mammana, C., & Gardini, L. (2000). Multistability and cyclic attractors in duopoly games. *Chaos Solitons Fractals, 11*, 543–564.

Bischi, G. -I., Chiarella, C., Kopel, M., & Szidarovszky, F. (2010). *Nonlinear oligopolies: Stability and bifurcations*. New York: Springer.

Cánovas, J. S., & Puu, T. (2010). On the dynamics of a piecewise linear oligopoly model with capacity limits and reinvestment periods. In T. Puu, & A. Panchuk (Eds.), *Nonlinear economic dynamics* (pp. 219–237). New York: Nova Science Publishers, Inc.

Cournot, A. (1838). *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette.

Edgeworth, F.Y. (1897). La teoria pura del monopolio. *Giornale degli economisti, 15*, 13–31.

Frisch, R. (1965). *Theory of production*. Dordrecht: D. Reidel Publishing Company.

Heathfield, F. H., & Wibe, S. (1987). *An introduction to cost and production functions*. London: MacMillan.

Kopel, M., & Szidarovszky, F. (2006). Resource dynamics under partial cooperation in an oligopoly. *Journal of Optimization Theory and Applications, 128* (2), 393–410.

Milnor, J. (1998). On the concept of attractor. *Communications in Mathematical Physics, 99*, 177–195.

Palander, T. (1939). Konkurrens och marknadsjämvikt vid duopol och oligopol. *Ekonomisk Tidskrift, 41*, 124–145, 222–250.

Panchuk, A., & Puu, T. (2009). Stability in a non-autonomous iterative system: An application to oligopoly. *Computer & Mathematics with Applications, 58*(10), 2022–2034.

Panchuk, A., & Puu, T. (2010). Industry dynamics, stability of Cournot equilibrium, and renewal of capital. In T. Puu, & A. Panchuk (Eds.), *Nonlinear economic dynamics* (pp. 239–254). New York: Nova Science Publishers, Inc.

Puu, T. (2005). Layout of a new industry: From oligopoly to competition. *Pure and Applied Mathematics, 16*, 475–492.

Puu, T. (2008). On the stability of Cournot equilibrium when the number of competitors increases. *Journal of Economic Behavior Organization, 66*, 445–456.

Puu, T., & Ruíz Marín, M. (2006). The dynamics of a triopoly Cournot game when the competitors operate under capacity constraints. *Chaos Solitons Fractals, 28*, 403–413.

Puu, T., & Sushko, I. (Eds.), (2002). *Oligopoly and complex dynamics: Models and tools*. New York: Springer.

Szidarovszky, F., & Okuguchi, K. (1998). An oligopoly model of commercial fishing. *Seoul Journal of Economics, 11*, 321–330.

Theocharis, R. D. (1959). On the stability of the Cournot solution on the oligopoly problem. *Review of Economic Studies, 27*, 133–134.

Tramontana, F., Gardini, L., & Puu, T. (2010). Global bifurcations in a piecewise-smooth Cournot duopoly game. *Chaos Solitons Fractals, 43*(1–2), 15–24.

Tramontana, F., Gardini, L., & Puu, T. (2011). Mathematical properties of a discontinuous Cournot Stackelberg model. *Chaos Solitons Fractals, 44*, 58–70.

R&D Networks

Gian Italo Bischi and Fabio Lamantia

Abstract The aim of this chapter is two-fold. It first provides a (non exhaustive) overview of the literature concerning oligopoly models where firms produce homogeneous goods and share R&D cost-reducing results through bilateral agreements and/or involuntary spillovers of knowledge. These models are expressed by the formalism of networks (i.e. theory of graphs) where firms represent nodes and agreements for R&D share represent arcs (or links). We describe several models and corresponding theoretical results, as well as some of their practical implications in industrial organization. The second part of the chapter describes a recent dynamic two-stage model of oligopolies with both R&D collaboration ties and spillover effects, where firms adaptively decide their R&D efforts myopically, through a boundedly rational process based on a local estimate of marginal profits. Moreover, a possible dynamic representation of knowledge accumulation with obsolescence is given.

1 Introduction

When firms compete for the production of high tech goods to be sold in a global market, their efforts are mainly devoted to gain knowledge in order to adopt new technologies and improve production standards. In many cases, such efforts can be identified with expenditures in *Research and Development* (*R&D*) activities, intended in a broader meaning ranging from basic notions to information about new technologies and skills in their adoption and employment. The overall result of R&D can be roughly summarized as a reduction of effective production costs (*process*

G.I. Bischi
Department of Economics, Society, Politics, University of Urbino "Carlo Bo", Via A. Saffi 42, Urbino, Italy
e-mail: gian.bischi@uniurb.it

F. Lamantia (✉)
Department of Economics, Statistics and Finance, University of Calabria, Via P. Bucci 0/1C, 87036 Rende (CS), Italy
e-mail: fabio.lamantia@unical.it

P. Commendatore et al. (eds.), *Complexity and Geographical Economics*,
Dynamic Modeling and Econometrics in Economics and Finance 19,
DOI 10.1007/978-3-319-12805-4_11

innovation) or creation of new/improved products (*product innovation*). Another important feature of knowledge gained by R&D activities is connected with the fact that it may spill from one firm to its competitors, due to the difficulty of protecting intellectual properties (see e.g. Spence 1984; D'Aspremont and Jacquemin 1988; Bischi and Lamantia 2002, just to cite a few). This unavoidable involuntary leakage of knowledge, whose causes may range from former employees who are hired by rival firms to industrial espionage, is often paralleled (and partially avoided) by voluntary agreements among rival firms to share R&D results and new technologies, for example through common training of workers and managers. In other words, sometimes firms which are competitors in the market behave as cooperators in the stage of developing research activities, finalization and installation of new cost-reducing technologies.

As a matter of fact, many empirical studies show that partnerships among firms have significantly increased in recent years (see e.g. Gauvin 1995; Goyal and Moraga-Gonzales 2001), and often this partnership is in terms of bilateral agreement for sharing information on R&D results and technological collaboration. The structure of bilateral partnerships and knowledge share (both voluntary and involuntary) among firms, can usefully be represented by the formalism of networks, where single firms are the nodes and knowledge share are the edges (or links) of the net.

In the recent years, the economic literature has been extensively focused on networks. In this survey, we review some of these recent models in the field of industrial competition between firms organized in Research and Development (R&D) networks. In most models, firms choose to collaborate in R&D in order to lowering costs of production as well as improving product quality.

The importance of R&D networks is well explained in Goyal and Joshi (2003), where different structures of bilateral collaborative links in firms' networks are described, as well as in Cowan and Jonard (2004), where the relationship between the network architecture and knowledge transmission is explored, mainly by numerical analysis. A well-known contribution on R&D networks is Goyal and Moraga-Gonzales (2001), which addresses the issue of collaborative ties formation starting from symmetric cases. For an extensive survey of the network literature we refer to Vega-Redondo (2007) and Goyal (2007).

As remarked in Goyal and Moraga-Gonzales (2001), collaborative relationships are often *nonexclusive*, i.e. firm i and j, j and k can have ties but i and k do not have any collaborative link. This is the main motivation for setting up oligopoly models with R&D networks structures. Indeed, traditional models of oligopoly are centered on markets and neglect the presence of such R&D networks.

In oligopolies with R&D networks, agents have several sources of strategic interdependence. First of all, oligopolists are strategically related on the demand side, as they all operate in the same market or in dependent markets. Moreover, with nonexclusive ties, also network externalities arise, as each firm's payoff is influenced by the R&D efforts by neighbors ("local" effects), non-neighbors ("global" effects) and by the whole set of connections in the industry. Of course, as remarked above, even without any agreement, knowledge may spill from one firm to its

competitors due to the difficulty of protecting intellectual properties (see Spence 1984; D'Aspremont and Jacquemin 1988; Bischi and Lamantia 2002). This clearly introduces another externality between agents and a free-riding dilemma into firms' relationships, so that a trade-off between partial (and often asymmetric) involuntary spillovers, and complete (symmetric) information share, associated with bilateral agreements, may arise.

The literature on R&D networks can be regarded as a generalization of the classic literature on R&D models, with firms linked in a network to share knowledge. For this reason, here we briefly review the basic models of oligopoly with R&D activities. The related issue of strategic location choice and R&D activities is surveyed in Kopel et al. (2015).

In this field, the seminal paper (D'Aspremont and Jacquemin 1988) proposes a two-stage game, where in the first stage two identical firms optimize their investment in cost-reducing R&D, with possible R&D spillover from the rival; then, in the second stage, firms compete in a homogeneous Cournot duopoly game. In particular, D'Aspremont and Jacquemin (1988) compare the equilibrium outcomes in two cases: in one the duopolists compete in both stages of the game; in another, firms choose R&D efforts to maximize joint profits in the first stage of the game while remaining competitors in the second stage. D'Aspremont and Jacquemin (1988) find that for high spillovers it is beneficial for firms as well as consumers to have cooperation in the R&D stage. However, the results are not so clear cut in the case of low spillovers, since a conflict between private and collective interests arise. Following this line of research, Kamien et al. (1992) proposed four different models, again in the form of two-stage games, where firms decide their effective R&D cost-reducing investments with or without formation of research joint ventures, and then they are engaged in either Cournot or Bertrand competition. All these models admit that research results are subject to various degrees of spillovers, and they conclude that the formation of research joint ventures, associated with competition in the product output, is the most desirable policy because it leads to higher profits and lower product prices. A dynamic version of the static game examined in D'Aspremont and Jacquemin (1988) has been recently proposed by Cellini and Lambertini (2009), in the form of a differential game, where it is shown that R&D cartelization dominates competition. Differently from D'Aspremont and Jacquemin (1988), Cellini and Lambertini (2009) show that taking into account investment smoothing, R&D cooperation is preferable to noncooperative behavior from a social as well as private point of view for any spillover level. Related issues in a difference game set-up are analyzed in Petit and Tolwinski (1996, 1999). Along the same line of research, see also Qiu (1997), Suzumura (1992),Amir et al. (1989) and Katz (1986). For more recent work, see Leahy and Neary (1997) and the references cited therein.

With respect to the research questions, the literature mainly address the following ones (see e.g. Goyal and Moraga-Gonzales 2001; Goyal and Joshi 2003; Bischi and Lamantia 2012a,b): Does a more connected firm exert a higher profit and earn a larger payoff with respect to a less connected firm? Is an increase in the collaboration level beneficial for all the firms? What are the effects of collaborative activity

on individual R&D and industry performance? What are the incentives of firms to collaborate and what is the architecture of "incentive-compatible" networks? Are individual incentives to collaborate adequate from a social welfare point of view? When is a network in equilibrium? Under which circumstances (level of collaboration, degree of knowledge spillovers, etc.) is an equilibrium stable, and how can it lose stability? How do the exogenous structural properties of competing networks influence aggregate outcomes? What are the influences of the degree of collaboration and involuntary spillovers on profits and, more generally, on overall efficiency?

This chapter is divided into two main sections: Sect. 2 gives a (non-exhaustive) overview about recent models and results on oligopoly models with firms arranged in R&D networks. While focusing on selected papers, we overview, in Sects. 2.2 and 2.3 respectively, models where the decision to establish a link rests with a single firm or a couple of firms. Then we briefly describe models with knowledge dynamics in Sect. 2.4. Section 3 explains a recent dynamic two-stage model of oligopolies with both R&D collaboration ties and spillover effects, with a boundedly rational adaptive process of R&D efforts. A possible dynamic representation of knowledge accumulation with obsolescence is then described in Sect. 3.3; Sect. 4 concludes.

2 Review of Oligopoly Models with R&D Networks

2.1 General Overview

The research on networks in industrial organization has become popular in the last two decades, and many papers have been devoted to a stylized description of partnerships among firms often in terms of bilateral agreement for sharing information on R&D results and technological collaboration (see e.g. Gauvin 1995; Goyal and Moraga-Gonzales 2001). Certain structural features of collaborative activity are typical in high-tech industries, such as biotechnology, pharmaceutical, chemical and computer industries, as pointed out empirically in Delapierre and Mytelka (1998), Hagedoorn and Schakenraad (1990) and Hagedoorn (2002).

As we shall see in this review, the distinctive features of R&D network models include the following ones: The assumption that the relationships between firms are non-exclusive (see Goyal and Moraga-Gonzales 2001; Bischi and Lamantia 2012a,b) and firms have the attitude to collaborate with other firms in the same market (Goyal and Moraga-Gonzales 2001; Goyal and Joshi 2003; Bischi and Lamantia 2012a,b; Bala and Goyal 2000); the nature of market competition (e.g. Cournot or Bertrand), see in particular the analysis in Billand and Bravard (2004); the cost of forming a link (Billand and Bravard 2004; Goyal and Joshi 2003; König et al. 2012); the architecture of the network (Cowan and Jonard 2004; Goyal and Moraga-Gonzales 2001; König et al. 2012; Meagher and Rogers 2004; Bischi and Lamantia 2012a,b); the efficiency and the stability of the networks (König et al.

2012; Westbrock 2010; Goyal and Moraga-Gonzales 2001; Jackson and Wolinsky 1996; Bloch and Jackson 2006); the direction of the information flow, which can be one-sided (see Billand and Bravard 2004) or two-sided (Goyal and Moraga-Gonzales 2001; Goyal and Joshi 2003; Bischi and Lamantia 2012a,b); the decisions to establish a link, which can depend on a single firm (Bala and Goyal 2000; Billand and Bravard 2004) or on couples of firms (Goyal and Moraga-Gonzales 2001; Goyal and Joshi 2003; Westbrock 2010; Bischi and Lamantia 2012a,b).

Formally, an R&D network can be defined through a graph $G = (N, E)$, where $N = \{1, 2, \ldots, n\}, n \geq 3$ is the set of nodes (firms) and $E \subseteq N \times N$ is a set of edges, that is, for any pair of nodes $i, j \in N$, the pair $(i, j) \in E$ if and only if i is connected to j. A useful representation of a network is obtained through the $n \times n$ adjacency matrix, whose ij element $g_{ij} \in \{0, 1\}$ assumes the value $g_{ij} = 1$ if i has a link with j, i.e. if $(i, j) \in E$, and $g_{ij} = 0$ otherwise. In most cases, the graph representing the R&D network is *undirected*, so that if $(i, j) \in E$ then also $(j, i) \in E$. This property clearly models situations where firms' R&D agreement are bilateral, with a two-way (symmetric) flow of information. In some papers (see Billand and Bravard 2004; Bala and Goyal 2000), the previous property does not necessarily hold and the underlying network is *directed*, in order to model cases where firm i can access to j's information but not vice versa, so that the flow of information is one-way (asymmetric). Another property often employed to simplify the analysis is that of *symmetry* (see Goyal and Moraga-Gonzales 2001; Bischi and Lamantia 2012a,b), i.e. that any firm in the network has the same number of collaborative ties, which represents the *level of collaborative activity* within the network. The importance of R&D networks is well explained in Goyal and Joshi (2003), where different structures of bilateral collaborative links in firms' networks are described, as well as in Cowan and Jonard (2004), where the relationship between the network architecture and knowledge transmission is explored, mainly by numerical analysis.

A central issue is the concept of an equilibrium network, that is a given network configuration where no firm has an incentive to break a link or links she has already set or to establish new links with other firms present on the market, holding the links of the other firms to be constant. In Jackson and Wolinsky (1996) it is stated that a link between two agents requires that both of them agree on the link (e.g. by making joint investments), therefore, the most common notions of stable networks rest on pairwise incentive compatibility, which lead to the notion of two-sided link formation. This concept is in particular employed in Goyal and Moraga-Gonzales (2001) and Bischi and Lamantia (2010, 2012a,b), see also Bloch and Jackson (2006) for a clear review of the most known notions of network stability employed in the literature.

In the next subsections we illustrate network models where the decision to establish links depends either on a single firm or on couples of firms. For other interesting strategic models of network formation we refer the reader to Aumann and Myerson (1989), Boorman (1975), Dutta et al. (1995), Jackson and Wolinsky (1996) and Kranton and Minehart (2001). With respect to the problem of coalition

formation in games, the main references are Bloch (1995), Kalai et al. (1979), Ray and Vohra (1997), Yi (1997) and the survey (Bloch 1997).

2.2 *Network Formation Models Through Individual Decisions*

In this subsection we focus on network formation models where the decision to establish links and the relative cost of forming and maintaining them are individual. In these models, the flow of information can be one-way, as in Billand and Bravard (2004), or two-ways, as in Bala and Goyal (2000) that consider both cases. In other words, the (R&D) network can be directed or undirected. Since individual agents decide whether establishing a link or not, the process of network formation can be modeled as a noncooperative game.

The impact of the nature of market competition (Cournot or Bertrand) in a directed network and the cost of forming a link is analyzed in Billand and Bravard (2004). In this paper, firms can establish one-sided flows of information and obtain positive cost externalities through them. The analysis is centered on firms' incentive to establish links and shows that the formation of equilibrium networks depend on the nature of market competition (Cournot or Bertrand) and on the cost for forming a link. In fact, in a Cournot oligopoly, if the cost of a link is low enough, the unique equilibrium network is the complete one, in which each firm picks up externalities of all other firms, whereas if the cost of setting links is high enough, the unique equilibrium network is empty, i.e. no firm gets information from its competitors (similarly to what is reported in Goyal and Joshi 2003). However, differently from Goyal and Joshi (2003), levels of cost of setting links exist such that, in equilibrium networks, some firms pick up externalities of all other firms while the others obtain no resources from their competitors. In the case of Bertrand oligopoly, the empty network still remains the unique equilibrium network when the cost of setting links is high enough. Otherwise, when the cost of setting a link is not too high, in the resulting equilibrium networks only one firm gets externalities from all competitors while the others obtain no externality from their competitors (inward-pointing star network). Given the one-way flow of information in the model, this implies that the unique firm obtaining externalities of the others has the lowest cost. This firm can set a price greater than its average cost and obtain positive profits. Thus, by contrast to the Cournot model, Billand and Bravard (2004) show that in the Bertrand model there exist conditions guaranteeing that some network architectures are in equilibrium, but they are not efficient. This points out that a conflict may arise between stability and efficiency in directed information networks.

In Bala and Goyal (2000), one-sided link-formation is considered, where an individual agent can form links with others by incurring some costs and each agent is a source of benefits that others can exploit via the formation of costly pairwise links. A link with another agent allows access to the benefits available to the latter via his own links. Thus individual links generate externalities whose value depends on the order of connections along chains of links. A distinctive aspect stressed by Bala and

Goyal (2000) is that the costs of link formation are incurred only by the person who initiates the link. Thus, an agent's strategy is a specification of the set of agents with whom he forms links. The links formed by agents define a network, and the focus is on benefits that are nonrival. Then both one-way and two-way flow of benefits are studied. In the former case, the link that agent i forms with agent j yields benefits solely to agent i, while in the latter case, the benefits accrue to both agents. In the benchmark model, the benefit flow across persons is assumed to be frictionless: if an agent i is linked with some other agent j via a sequence of intermediaries, then the benefit that i derives from j is insensitive to the number of intermediaries. Moreover, they allow for a general class of individual payoff functions: the payoff is strictly increasing in the number of other people accessed directly or indirectly and strictly decreasing in the number of links formed. The main result of the paper is that networks where all agents play a Nash equilibrium ("Nash" networks) are either connected or empty.

2.3 Network Formation Through Pairwise Collaborative Links

In this subsection, we consider the network formation problem with links established through decisions of couple of firms. For these models, the flow of information is always two-ways, i.e. the underlying networks are always directed.

An important contribution to the study of group formation and cooperation in (Cournot) oligopolies is Goyal and Moraga-Gonzales (2001), inspired by works on strategic models of network formation such as Aumann and Myerson (1989), Bala and Goyal (2000), Dutta et al. (1995), Goyal and Joshi (2003), Jackson and Wolinsky (1996), Kranton and Minehart (2001). It studies the incentives for collaboration between horizontally related firms with two-sided link formation. In their model, a firm invests in a cost-reducing R&D technology, and when firms collaborate their individual R&D efforts lower also their partners' costs. The impact of these cost-reducing bilateral agreements is related to the nature of market rivalry between the firms, i.e. whether the firms operate in independent markets or not. By forming collaborations, however, firms alter the competitive position of different firms thus influencing market structure and performance, and Goyal and Moraga-Gonzales (2001) investigate whether R&D activities depend on the nature of market competition. In particular, they consider an oligopoly with (ex ante) identical firms. Prior to market interaction, each firm has an opportunity to form pairwise collaborative links with other ones in order to share R&D knowledge about cost-reducing technologies. The collection of pairwise links between the firms defines a network of collaboration with nonexclusive ties. Given such a collaboration network, firms choose a (costly) level of effort in R&D unilaterally, aimed at reducing their own production costs. Then the level of effort by different firms and the structure of the R&D network define the effective costs of the different firms in the market. Given these costs, firms compete in the market by setting quantities (i.e. Cournot competition). Concerning the types of market interaction,

Goyal and Moraga-Gonzales (2001) consider two cases: in the first, firms operate in independent markets, while in the second case, they compete in a homogeneous-product market. Analytical results are obtained with symmetric networks, i.e., networks in which all firms maintain the same number of collaborative ties. For such networks, the level of collaborative activity is measured in terms of the number of ties of a typical firm. Their first result pertains to the relationship between the level of collaboration and individual R&D: if firms compete in a homogeneous-product market, individual R&D effort is declining in the level of collaborative activity. In contrast, when firms operate in independent markets, individual R&D effort increases monotonically in the level of collaborative activity. Thus, in the former case individual R&D attains its minimum, whereas in the latter case it attains its maximum under the complete network, i.e., a network in which every pair of firms has a collaborative agreement. This contrast highlights the influence of the competition effect. Then they examine the level of cost reduction under different levels of collaboration. For any given level of R&D effort, adding a collaboration link leads to lower costs for all firms. However, in a homogeneous-product market a higher level of collaborative activity lowers individual research efforts. Thus they compare the relative magnitude of these two effects. The main finding is that in a homogeneous-product setting, the level of cost reduction is initially increasing and then decreasing in the level of collaborative activity, i.e., it is non-monotonic with respect to the number of collaborations. By contrast, when firms operate in independent markets, individual R&D effort and the cost reduction is maximal under the complete network.

Another interesting result in Goyal and Moraga-Gonzales (2001) concerns the incentives of firms to form collaborative alliances. Irrespective of the degree of market competition, firms have an incentive to form links, hence the empty network is never incentive compatible. Furthermore, the incentives to form collaborations are quite large in both settings: the complete network is a strategically stable network, irrespective of the market setting. Concerning the aggregate industry performances, if firms compete in a homogeneous-product market, industry performance is highest (both in terms of aggregate profits and social welfare) when firms have an intermediate level of homogeneous degree of collaboration with other firms. This means that both the empty and the complete networks are dominated by intermediate levels of collaboration. Thus, under market rivalry, the incentives of firms to form R&D collaboration links may be excessive both from an industry-profit-maximizing perspective as well as from a social welfare point of view. By contrast, if firms interact in independent markets, aggregate industry profits as well as social welfare are highest under the complete network.

These results on welfare and profit properties of collaboration could provide an explanation for why a large number of strategic alliances are unstable or are terminated early, and they also help explain why some alliances work well. In highly competitive environments, in order to get higher profits firms may "collectively" prefer not to form many collaborative ties. However, a pair of individual firms gain competitive advantage over the rivals by forming a collaboration and thus increase their profits. This implies that firms may individually have incentives to form too

many links, thus leading to poor overall performance. However, this dilemma does not appear in the case of independent markets, because firms' R&D efforts are not declining with the number of links. In the first part of their analysis, Goyal and Moraga-Gonzales (2001) restrict the analysis to symmetric network structures, where every firm is ex-post in a similar situation. They next examine the role of collaborations in generating such competitive advantages and their influence on market structure and industry performance. This motivates an examination of asymmetric networks. A general analysis of asymmetric networks, however, turns out to be very complicated; therefore Goyal and Moraga-Gonzales (2001) work out only an example with three firms, completely characterizing its solution. In this setting there are four possible network architectures: the complete network, the star network, the partially connected network, and the empty network. Asymmetric networks such as the star or the partially connected network perform quite well from the social as well as the private point of view. Indeed, the star network always dominates the complete network from both perspectives; moreover, for some parameter values, the star is industry-profit maximizing. In addition, asymmetric forms of collaboration may alter the market structure by causing large disparities between firms, or even leading to the exit of firms, and that this is not necessarily detrimental from a social standpoint. Indeed, under certain circumstances, the partially connected network is strategically stable, and both industry-profit and social-welfare maximizing.

Following a similar stream, Goyal and Joshi (2003) studies networks of collaboration between oligopolistic firms and shows how market competition is crucial in defining R&D network structures. Differently from Goyal and Moraga-Gonzales (2001), where costs of forming links are taken to be negligible but firms decide independently on a level of R&D, Goyal and Joshi (2003) assume that a collaboration link between two firms involves a fixed cost and leads to an exogenously specified reduction in marginal cost of production, and consider an oligopoly setting in which firms form pairwise collaborative links with other firms. These pairwise links involve a commitment of resources on the part of the collaborating firms and yield lower costs of production for firms which form the link. The collection of pairwise links defines a collaboration network and induces a distribution of costs across the firms in the industry. Given these costs, firms then compete in the market. Their analysis focuses on how the costs of forming links affect the architecture of strategically stable networks. In this model, a firm can form collaboration relations with other firms without seeking prior permission of current collaborators. This has important strategic effects and requires novel methods of analysis. Firstly Goyal and Joshi (2003) study an example in which the cost for forming a link is negligible, and in this case a complete characterization of strategically stable and socially efficient networks is provided. Under quantity competition, the complete network is the unique stable and socially efficient network. With price competition the empty network is the unique stable network, while the efficient network is an inter-linked star, with two central firms. The results in Goyal and Joshi (2003) show that the nature of market competition has an important effect on the type of collaboration networks that arise and on the level of welfare. In addition, Goyal and Joshi (2003)

considers network formation under general market conditions. In a market with moderate competition, all firms make positive profits but firms with lower cost make larger profits. Thus quantity competition under homogeneous or differentiated demand, and price competition under differentiated demand, are special cases of this type of competition. In this case, every pair of firms with the same costs must be linked, and this implies that in the class of symmetric networks only the complete network can be stable. Then, the authors consider the case where all lowest cost firms make positive profits. Under aggressive competition, stable networks have the dominant group architecture, i.e. firms divide themselves into two groups, with one group containing at least three firms and having the feature that every pair of firms has a collaboration link, while the second group consists of isolated firms. When the link formation has a high cost, the analysis in Goyal and Joshi (2003) is carried on with the linear demand Cournot model. In this case, it holds that firms have increasing returns from links and, for a given class of parameters, a stable network has the dominant group architecture, with the size of this group being sensitive to the cost of forming links. An interesting aspect of the analysis is a non-monotonicity in the sustainable size of the dominant group as the costs of forming links increase. For small costs of forming links, only a large dominant group is stable, at intermediate costs large as well as small groups can be stable, while for large costs only a medium sized group is stable. The property of increasing returns from link formation suggests that a firm with many links may have an incentive to induce a firm with few links to form a collaboration relationship by offering to subsidize its costs of link formation. This motivates an examination of stable networks when transfers are allowed between firms. In this case, a stable network has the dominant group architecture or is an interlinked star.

The social efficient networks of Goyal and Joshi (2003) are investigated by Westbrock (2010) as well, where it is shown that asymmetric networks are typically efficient. In fact, Westbrock (2010) demonstrates that the social efficient network may be the empty network or the complete network. Otherwise, if neither of these are efficient, then efficient networks have a dominant group or an interlinked star architecture, i.e. they have an asymmetric architecture. In addition, Westbrock (2010) analyzes the density and the degree variance in the efficient network, by expressing social welfare as an additive function of the density and the degree variance of the network.

2.4 Knowledge Dynamics

A central issue in R&D network models is related to the dynamics of the flow of knowledge between linked firms. In Cowan and Jonard (2004) knowledge diffusion is modeled as a dynamic process in which agents exchange different types of R&D results. Agents are located on a network and are directly connected with a small number of other agents, and each agent repeatedly meets those who are directly connected with her and trade if mutually profitable trades exist. In this way

knowledge diffuses throughout the economy. The emphasis is on the relationship between network architecture and diffusion performance, and in particular they model, at one extreme, a network in which every agent is connected to the nearest neighbors, and at the other extreme a network with each agent being connected to, on average, n randomly chosen agents. Then they consider the set of structures that fall between the two opposite extreme cases. The main finding is that the performance of the system clearly exhibits small world' properties, as the steady-state level of average knowledge is maximal when the structure is a small world (that is, when most connections are local, but roughly 10 % of them are long distance).

König et al. (2012) is a very interesting contribution on the field, where the growth of knowledge within a firm follows a continuous time dynamical system. Based on other papers of the same authors (see, in particular, König et al. 2011), König et al. (2012) first characterize the topology of the efficient structure of links for varying marginal cost of collaborations. In fact, when the marginal costs of collaboration are low, the trade-off between number of walks and costs of direct connections is loose, and the complete network is efficient. For intermediate values of marginal collaboration costs, efficient graphs are still connected but it becomes efficient to save on the number of direct collaborations and to create a high number of walks by concentrating connections among a small number of firms. More precisely, efficient graphs are nested split graphs, i.e. the neighborhood of each node is contained in the neighborhood of the next higher degree nodes. This implies a strong hierarchical degree structure of the network. As costs of collaboration become high, asymmetric efficient graphs can become disconnected and then empty. Following Jackson and Wolinsky (1996), the emergence of pairwise stable structures is also considered in König et al. (2012): the complete graph is stable if marginal costs of collaboration are low and industry size is small. In large industries, the set of disconnected cliques of homogeneous size and the star are stable. Both structures can be stable for the same values of industry size and collaboration costs. In addition, the size of stable cliques decreases with marginal collaboration costs. The empty graph is stable for very large marginal costs and any industry size. Finally, the relationship between stability and efficiency is studied. Here it is shown that efficiency is reached in industries of small size and low cost of collaboration. As industry size grows, however, firms have lower incentives to form collaborations, and a divergence between stability and efficiency emerges.

3 A Dynamic Model of R&D Efforts Along Networks

3.1 Model Setup

In this section, we briefly review a discrete-time dynamic model recently proposed in Bischi and Lamantia (2012a,b) to simulate the time evolution of R&D efforts of firms arranged in subnetworks of R&D collaboration ties in the presence of

inter- and intra-subnetwork involuntary spillovers of R&D results. The model focuses on the dynamics of R&D efforts when the network structure is given, thus extending Goyal and Moraga-Gonzales (2001), where effort levels are always assumed in equilibrium. Since more than one R&D network can be present, the model can be employed to analyze oligopolistic competition with regional clusters of firms sharing R&D results and global competition in a unique market.

In detail, a repeated two-stage oligopoly is proposed where N ex-ante identical firms are subdivided into one or more symmetric groups (denoted as "subnetworks") characterized by a given degree of homogenous connections. In these subnetworks couples of firms (denoted as "neighbors") have bilateral ties to share R&D results. Hence, as usual in R&D networks, effective costs of each firm include positive cost externalities not only as a result of its own cost-reducing R&D (expensive) efforts, but also the advantages of their neighbors' efforts. Moreover, a firm can receive, for free, two types of knowledge spillovers: (i) internal (from non-neighbors inside its subnetwork) or (ii) external (from firms outside its subnetwork). In the model, each discrete time step is ideally subdivided into:

- A precompetitive stage, where each firm selects cost-reducing R&D efforts in the direction of increasing individual profits, following positive marginal profits in the steepest ascent direction (so-called "gradient process");
- A (Cournot-)competitive stage, where each firm sets its "optimal" quantity, taking into account the current level of efforts of other firms and the networks' structures in the computation of its effective cost.

These disjoined subnetworks can be interpreted as different countries or industrial districts or groups of firms linked by ownership ties, characterized by different rules for partnership or patent protection or different abilities to take advantage from spillovers.

The model proposed describes an homogeneous-product oligopoly, where $N \geq 2$ quantity setting firms operate in a market characterized by a linear inverse demand function $p = a - bQ, a, b > 0$, Q being the total output in the market. These N firms are ex-ante partitioned into h groups, (called *subnetworks*). Two firms of the same subnetwork are neighbors if they have a direct link, i.e. they form a bilateral agreement for a complete sharing of R&D results. Two firms without a direct link are called non-neighbors. Each of these h subnetworks, say s_j, $j = 1, \ldots, h$, is formed by n_j firms, where of course $N = \sum_{j=1}^{h} n_j$. Each subnetwork s_j is assumed symmetric of degree k_j, with $0 \leq k_j \leq n_j - 1$, i.e. every firm in s_j has the same number of collaborative ties k_j, a parameter that represents the level of connectivity (or collaborative attitude) of subnetwork s_j.

Each firm decides its R&D effort, whose cost-reducing effects are totally shared with neighbors; moreover, R&D results within a subnetwork can spill over for free to non-neighbors inside the same subnetwork s_j (internal spillovers) as well as to "rival" subnetworks s_k with $k \neq j$ (external spillovers). Assuming a linear cost

function $c_i q_i$ for a firm i that produces q_i, a representative firm in subnetwork s_j has a marginal cost c_i of the form

$$c_i(s_j) = c - e_i - k_j e_{l_i} - \beta_j e_{l_{-i}} \left[(n_j - 1) - k_j \right] - \beta_{-j} \sum_{m \notin s_j} e_m \tag{1}$$

where c is the marginal cost without R&D efforts (equal for all firms), e_i represents R&D effort of firm i, $k_j e_{l_i}$ represents the total effort exerted by firms with whom i is directly linked in s_j, $\beta_j \in [0, 1)$ are related to spillovers with non-neighbors in network s_j, and $\beta_{-j} \in [0, 1)$ regulate external spillovers, i.e. originating from non-neighbors out of s_j towards firm i. We denoted by l_i and l_{-i} a representative firm in s_j firm i is, respectively, linked and not linked. Hence, the two last two terms in (1) are, respectively, the spilled effort by l_{-i} and m, i.e. representative non-neighbors inside s_j and outside s_j.

In any case, all N firms are rivals in the market place (see D'Aspremont and Jacquemin 1988; Goyal and Moraga-Gonzales 2001), and they calculate their optimal quantity by solving individual profit maximization problems. Then, given optimal quantities as function of efforts by backward induction, they assess R&D efforts to increase their individual profits; due to R&D networks of collaboration and spillovers, each firm calculates these cost-reducing efforts taking into account not only the network structure it belongs to (number of firms in the network and number of neighbors) but also the whole cost structure of the industry. Each oligopolist i in subnetwork s_j maximizes a profit function of the form

$$\pi_i(s_j) = \left\{ a - b \left[q_i(s_j) + \sum_{p \neq i} q_p \right] - c_i(s_j) \right\} q_i(s_j) - \gamma e_i^2(s_j) \tag{2}$$

where $q_i(s_j)$ and $e_i(s_j)$ are, respectively, the quantity and the R&D effort by agent i in subnetwork s_j, and γe_i^2, $\gamma > 0$, is the cost of effort.

The optimal quantity of firm i in subnetwork s_j is

$$q_i(s_j) = \frac{a - N c_i(s_j) + \sum_{p \neq i} c_p}{b\,(1 + N)} \tag{3}$$

with corresponding optimal profit

$$\pi_i(s_j) = \left[\frac{a - N c_i(e_i) + \sum_{p \neq i} c_p}{\sqrt{b}\,(1 + N)} \right]^2 - \gamma e_i^2 \tag{4}$$

Given this setting, each firm tries to maximize its individual profit with respect to its own R&D efforts. Substituting the cost functions of representative agents in each subnetwork, the optimal profit for firm i in subnetwork s_j (4) can be rewritten as a quadratic functions of efforts only, where the sum of marginal costs of all firms but i appearing in (3) and (4) is given by

$$\sum_{p\neq i} c_p = k_j c_{l_i}(s_j) + \left(n_j - 1 - k_j\right) c_{l_{-i}}(s_j) + \sum_{w=1, w\neq i}^{h} n_w c_v(s_w) \tag{5}$$

The last term in (5) represents the marginal cost of all firms that are not in the same subnetwork of firm i.

By standard arguments, it is possible to establish the existence of a Nash equilibrium. Under concavity of payoffs in own strategies, the FOCs $\frac{\partial \pi_i}{\partial e_i} = 0, i = 1, \ldots, h$ are necessary and sufficient for an interior optimum E^*. As subnetworks are symmetric, all firms belonging to the same subnetwork exert the same effort, (i.e. $e_l = e_{l_{-i}} = e_i$). When all firms start the game exactly at the Nash equilibrium, then there is no unilateral incentive to deviate and the model reduces to a one shot game. Otherwise a dynamic mechanism for updating R&D effort over time must be defined. Before doing so, we briefly analyze the main relationships between marginal profits and effort in the proposed multi-network competition.

Due to the complex network structure of R&D collaborations and spillover externalities, it is unlikely that agents are able to play the Nash equilibrium strategy in one shot. Consequently, firms are considered as boundedly rational, i.e. they do not have complete information about the strategic variables, and each player is assumed to care about immediate local profit maximization and adaptively adjust their efforts over time along the direction of the local (correct) estimate of expected marginal profits, according to the so called "gradient dynamics" (or "gradient process", see Arrow and Hurwicz 1960; Furth 1986; Flam 1993; Bischi and Naimzada 1999; Friedman and Abraham 2009)

$$e_j(t+1) = e_j(t) + \alpha_j(e_j)\frac{\partial \pi_j}{\partial e_j}; \quad j = 1, \ldots, h \tag{6}$$

where $e_j(t)$ represents the R&D effort at time period t of a representative firm belonging to the subnetwork s_j; α_j are positive functions that represent speeds of adjustment.

Notice that each firm uses naive (or static) expectations about other firms' effort choices, i.e. they believe that actions of other players in the next period will be the same as in the current period (see Goyal 2007). So, efforts at time t, which are observable by all agents, lead to the choice of next period R&D activities, through a repeated adaptive process typical of boundedly rational agents (6). It can easily be seen that the Nash equilibria, that fully rational (and informed) players are assumed to be able to select in one shot, are also equilibrium points for the boundedly rational

dynamic process (6). If such an equilibrium is stable, then we can say that the adaptive agents are able to learn, in the long run, how they can choose R&D efforts in an optimal way. However, as we shall see, these equilibria are not always stable under the gradient dynamics (6).

The focus is on the case of only two subnetworks s_1 and s_2 with n_1 and n_2 firms and connection degrees k_1 and k_2 respectively. Moreover, speeds of adjustment are assumed linear with $\alpha_j(e_j) = \alpha_j e_j$, i.e. the relative effort change $[e_j(t+1) - e_j(t)]/e_j(t)$ is posited to be proportional to the expected marginal profit.

Under these assumptions the dynamical system that describes the time evolution of the efforts chosen by the two representative firms is given by

$$e_i(t+1) = e_i(t) + \frac{\alpha_i e_i(t)}{b\left(1+n_i+n_j\right)^2}\left[A_i + B_i e_j(t) + C_i e_i(t)\right],\ i,j = 1,2; i \neq j \tag{7}$$

where:

$$\begin{aligned} A_i &= 2(a-c)\left[(n_i - k_i)(1-\beta_i) + \beta_i + n_j(1-\beta_{-j})\right] \\ B_i &= 2n_j\left[(1-\beta_i)(n_i-k_i) + \beta_i + n_j\left(1-\beta_{-j}\right)\right]\cdot \\ &\quad \cdot\left[-\beta_j\left(n_j - k_j - 1\right) + \beta_{-i}\left(n_j+1\right) - k_j - 1\right] \\ C_i &= 2\Big\{(-k_i + \beta_i(1+k_i-n_i) + N - \beta_{-j}n_j)\cdot \\ &\quad \cdot(1 + k_i + n_j + k_i n_j - \beta_{-j}n_i n_j - \beta_i(1+k_i-n_i)(1+n_j)) \\ &\quad -b\gamma\,(1+N)^2\Big\} \end{aligned} \tag{8}$$

with $N = n_1 + n_2$.

The dynamical model (7) always admits four equilibria. There are three boundary equilibria: $O = (0,0)$ and

$$E_1 = (-A_1/C_1, 0) \text{ and } E_2 = (0, -A_2/C_2), \tag{9}$$

located along the invariant coordinate axes, with nonzero coordinate strictly positive if and only if the corresponding profit function $\pi_i(e_i)$ is strictly concave. Moreover, provided that $C_1C_2 - B_1B_2 \neq 0$, a unique interior equilibrium

$$E^* = \left(\frac{A_2B_1 - A_1C_2}{C_1C_2 - B_1B_2}, \frac{A_1B_2 - A_2C_1}{C_1C_2 - B_1B_2}\right) \tag{10}$$

We observe that, in general, at a boundary equilibrium E_i only n_i agents invest in R&D even if N agents sell their product in the market. Equilibrium E^* is obtained by unilateral profit maximization by representative firms in the two subnetworks and

correspond to the Nash equilibrium solution described in the previous section. We characterize the equilibrium E^* in some benchmark cases in the following.

3.2 Main Results

The main analytical results in Bischi and Lamantia (2012a) are obtained through the study of the two-dimensional map (7), and in particular through the local stability analysis of the various equilibria of the model. A first result of Bischi and Lamantia (2012a) is that for a sufficiently low initial R&D level (no matter how low) there will be at least one network exerting a positive (and increasing with time) level of effort, i.e. it is always convenient, for at least one network, to invest in R&D.

Another analytical result in Bischi and Lamantia (2012a) shows that if the cost of effort γ is sufficiently high and only network i invest in R&D, then this network will tend to adopt a constant level of efforts provided that its reaction coefficient α_i or the aggregate parameters A_i are sufficiently small: we can relate "small" A_i values to small differences between the maximum selling price a and the maximum marginal cost coefficient c and/or to a great cost reduction within network i (due to many collaborative arrangements within network i or high internal spillovers). Otherwise instability can be present and R&D efforts inside network i are characterized by increasing levels of unpredictability.

When $E_i, i = 1, 2$, is stable also in the direction transverse to the coordinate axis e_i, then we observe a decrease in R&D efforts by network j till its effort goes to zero, and only network i will invest at the equilibrium level E_i.

Other analytical results in Bischi and Lamantia (2012a) are obtained for two opposite benchmark cases. In the first one, all firms compete in the market and possibly invest in R&D, but no ties are established among them (*empty network*). Thus, individual effort has never the effect to reduce costs to competitors, neither in form of agreements nor in form of involuntary spillovers. The second benchmark is, in some sense, the opposite case, represented by two competing networks that are fully connected, i.e. each firm completely shares its R&D cost-reducing results within its network (*complete networks*).

Due to the hypothesis of symmetry, the study with two competing subnetworks can be carried out analytically by studying a map of the plane, thus establishing two main results. The first one is related to transverse stability of boundary equilibria. Bischi and Lamantia (2012a) demonstrate that they are always unstable with empty subnetworks, but can be stable with complete subnetworks. This fact has an interesting interpretation in terms of dominant group architecture of the industry. The second result is a welfare analysis of equilibria for the benchmarks, showing that empty subnetworks are never efficient.

When the inner effort equilibrium E^* is positive and stable for the cases of an empty network and a complete network, it is interesting to carry out some consideration on welfare analysis at the positive equilibrium (see Qiu 1997 for the analysis in the Cournot game without network structure). The results for the case

with a single network (e.g. $n_2 = 0$) corresponds exactly to proposition 7 and 8 in Goyal and Moraga-Gonzales (2001). For sake of simplicity it is assumed $b = 1$.

In case of empty subnetworks, by making use of E^*, the equilibrium quantity and individual profits of a representative firm, with consumer surplus are given by

$$q^E = \frac{(a-c)\,\gamma\,(1+N)}{\gamma(1+N)^2 - N}; \pi^E = \frac{(a-c)^2 g\left(\gamma(1+N)^2 - N^2\right)}{\left(\gamma(1+N)^2 - N\right)^2}; \qquad (11)$$

$$CS^E = \frac{(a-c)^2\gamma^2 N^2(1+N)^2}{2\left[\gamma(1+N)^2 - N\right]^2}$$

where the index E stays for "Empty".

Correspondingly, for the case of complete subnetworks,[1] equilibrium quantity and individual profits of a representative firm with consumer surplus respectively are given by

$$q^C = \frac{(a-c)\,\gamma\,(1+2n)}{\gamma(1+2n)^2 - n\,(1+n)}; \pi^C = \frac{(a-c)^2\gamma\left(\gamma(1+2n)^2 - (1+n)^2\right)}{\left(n+n^2 - \gamma(1+2n)^2\right)^2};$$

$$CS^C = \frac{2(a-c)^2\gamma^2 n^2(1+2n)^2}{\left[n+n^2 - \gamma(1+2n)^2\right]^2} \qquad (12)$$

where the index C stays for "Complete". Moreover, in the following we denote the total welfare by $TW = N\pi + CS = \sum_{i=1}^{N}\pi_i + \frac{1}{2}\left(\sum_{i=1}^{N} q_i\right)^2$.

By direct comparison of the previous quantities it is possible to show that if $b = 1, n_1 = n_2 = n$ then

$$\pi^E < \pi^C; CS^E < CS^C; TW^E < TW^C.$$

Under these circumstances, it is preferable, both from a private and a social point of view, to have a competition between two complete networks than two empty ones. These results are also confirmed in the numerical simulations with random networks reported in Bischi and Lamantia (2012b). Of course, more efficient solutions for intermediate levels of collaboration activity in the subnetworks are not ruled out.

Summing up, the results in Bischi and Lamantia (2012a,b), when two networks with equal level of connectivity compete, spillovers that are internal to a network can reduce progressively the market share of the rival network. Furthermore, in cases where a network starts with the disadvantage of having less links than its competing network, internal spillovers can completely overturn the positions, with discontinuous changes in equilibrium investment levels and profits. This is due to the coexistence of different long run attractors, i.e. to coexisting fixed points in

[1] Note that $n_1 = n_2 = n$.

the space of efforts, each with its own basin of attraction, leading to a situation of *multistability*. With respect to external spillovers, we observe that when they are low, they provide substantial benefits to the network they are directed to, whereas above a threshold level, they turn out to be harmful for the receiving network. For all these cases we explored also the effects of these parameters on social indicators of performance.

In Bischi and Lamantia (2012b), the assumption on networks' symmetry is relaxed, showing how the framework of the paper can be adapted to perform numerical analysis with asymmetric networks, which are randomly generated.

3.3 Knowledge Accumulation in a R&D Network

When dealing with R&D efforts, whose result can be described as production of knowledge, memory effects should be considered for this immaterial form of capital, also denoted as "knowledge capital stock". In fact, knowledge gradually accumulates as a result of past efforts (see e.g. Nelson 1982; Cohen and Levinthal 1989, 1990; Confessore and Mancuso 2002), but it usually can be only partially "capitalized" because knowledge tends to depreciate itself as time goes on, according to the obsolescence rate of information and technology. In Bischi and Lamantia (2010), a general framework is proposed to describe an industry, evolving in discrete time, where firms invest in R&D efforts in order to increase their overall knowledge level, otherwise subject to obsolescence (see Bischi and Lamantia 2004; Petit and Tolwinski 1999; Petit et al. 2000). More precisely, time t investments in R&D by firm i, denoted by $E_i(t)$ increases the time t total (or accumulated) knowledge of firm i, which is given by

$$z_i(t) = \sum_{k=0}^{t} \rho^{t-k} E_i(k) = E_i(t) + \rho \sum_{k=0}^{t-1} \rho^{t-1-k} E_i(k) = E_i(t) + \rho z_i(t-1) \qquad (13)$$

where $\rho \in [0, 1)$ gives a measure of how rapidly information becomes obsolete. A similar formulation of knowledge accumulation was also employed in Bischi and Lamantia (2004), where an increment of total knowledge has a cost-reducing effect, as a firm i has a marginal cost function at time t of the form

$$c_i(t) = c_0 - c_0 f_i\left(z_i(t)\right) \qquad (14)$$

where c_0 is the marginal cost without R&D efforts (equal for all firms) and $f_i\left(z_i\right)$ is an R&D production function (see on this point Confessore and Mancuso 2002).

Moreover, firms that invest in R&D can establish bilateral agreement with other competitors and create an R&D network; this means that they can cooperate and share R&D results, even if they remain competitors on the marketplace, as proposed in D'Aspremont and Jacquemin (1988) and Goyal and Moraga-Gonzales (2001).

However, R&D investments are not entirely private, as a fraction of them can spill over for free to competitors (see also Audretsch and Feldman 1996; Bischi and Lamantia 2002; Bischi et al. 2003 on this point). In Bischi and Lamantia (2010), at each time period each firm has two different choices to make: the level of R&D effort to exert, and then the quantity to produce. At each time period firms are able to "solve" the problem of optimal production choice, according to the R&D efforts, by backward induction, similarly as Bischi and Lamantia (2012a). In this way, the model becomes a repeated two stage game, so that the choice of an R&D effort strategy is always followed by an "optimal" subsequent choice of quantities. Firms are homogeneous, so that also total knowledge is an homogeneous quantity within the industry. The network structure is fixed, so that firms have to decide the level of their R&D efforts and the quantity to produce within a given network structure. With respect to the effect of a change in the level of collaboration activity, firms' agreements for sharing R&D results are more alliances (long-term instances) than coalitions (short-lived instances), as it is often the case when agreements originate from Joint Ownership relationship (see Gauvin 1995). After describing the general framework of the model, a specific functional form for the R&D production function is proposed. Also in this case, a dynamic formulation of the discrete-time model is expressed in terms of gradient dynamics (see Flam 1993; Furth 1986; Bischi and Naimzada 1999; Szabó and Fáth 2007). As a consequence of the assumption on agents' homogeneity, this dynamic model is two-dimensional, with dynamic variables given by R&D effort and knowledge. In particular, the equilibria of the dynamic model are also equilibria for the corresponding static model. In this way the proposed dynamic model is useful for two different purposes: first, it describes the "out of equilibrium" decisions of a representative firm that engages the competition over time. Second, in case of convergence to an equilibrium point, the dynamic model is a numerical tool for finding these equilibria, that, in general, can not be found algebraically. In fact, analytical results on existence, uniqueness and stability of an R&D effort equilibrium are provided with a strict concave R&D production function, i.e. when the marginal productivity of knowledge is decreasing. When the R&D production function obeys the "law of variable proportions", some insights can be given by numerical methods. Both the cases of a concave R&D production function and a convex-concave one are explored. For the first case a unique equilibrium for R&D efforts exists with a corresponding stationary level of total knowledge of the industry. With a convex-concave production function, it is possible to observe a discontinuous transition from a positive equilibrium to absence of investments as spillovers are increased; this phenomenon shows hysteresis effects, so it can even be irreversible as spillovers are reduced back to the previous level. This suggests that it is important to protect intellectual properties to limit the disincentive to invest in R&D as a consequence of free rider behaviors. Moreover, the numerical experiments suggest that, both in the case of strictly concave or convex-concave production function, R&D efforts are decreasing as the number of links in the network is increased. Total knowledge appears to be always maximized for intermediate levels of collaboration activity.

However, not only the cost reduction effect of R&D activities but also the capacity to exploit spillovers (i.e. the "absorptive capacity", see Confessore and Mancuso 2002) should be assumed to depend on the accumulated knowledge. Moreover, it would be more reasonable to assume that internal spillovers are not constant, but their effects fade with the network distance among competitors.

4 Conclusion

In this chapter we have given a general (but not exhaustive) review of the literature on oligopoly models with R&D activities, with firms linked in a network according to bilateral agreements for sharing cost-reducing R&D results, i.e. knowledge and new technologies. The results given in this literature stream address several important issues concerning industrial organization, creation of industrial clusters, incentives to reach an optimal level of collaboration among firms in order to increase profits and overall efficiency, determination of stable networks' structures.

The possibility of a dynamic setup of oligopoly models along R&D networks has also been analyzed in this chapter, both from the point of view of literature overview and in a particular model based on a repeated two-stage game used to describe the competition between firms arranged in a set of R&D subnetworks, and acting non-cooperatively at the second stage, where quantities to sell are decided in order to maximize their individual profit, as in a Cournot oligopoly game. The long run outcome of such adaptive process may coincide with an underlying Nash equilibrium of the game. However, when several equilibria are present, as it occurs in the model proposed when inner and boundary equilibria coexist, an *equilibrium selection* problem arise, so it is crucial to analyze the path dependent dynamic transition toward an equilibrium, as the initial condition of the system plays a role in the long run behavior of the model. This point recalls Nash's concern about a possible evolutionary interpretation of the concept of Nash equilibrium.

A remarkable improvement of the models dealing with R&D networks can be obtained by assuming that knowledge gained through R&D efforts accumulates over time. Following Bischi and Lamantia (2004), a first step in this direction can be found in Bischi and Lamantia (2010), where an R&D network game with knowledge accumulation is analyzed. However, both the cost reduction effect and the capacity to exploit spillovers (i.e. the "absorptive capacity", see Confessore and Mancuso 2002) should be assumed to depend on the accumulated knowledge. Other issues that may be worth to be investigated include the dynamics of formation of emerging networks at different scales (regional, national and international), the location choice and formation of clusters (or industrial districts).

References

Amir, R., Evstigneev, I., & Wooders, J. (1989). Noncooperative versus cooperative R&D with endogenous spillover rates. *Games and Economic Behavior, 42*, 183–207.

Arrow, K. J., & Hurwicz, L. (1960). Stability of the gradient process in n-person games. *SIAM Journal on Applied Mathematics, 8*(2), 280–294.

Audretsch, D. B., & Feldman, M. P. (1996). R&D spillovers and the geography of innovation and production. *American Economic Review, 86*, 630–640.

Aumann, R., & Myerson, R. (1989). Endogenous formation of links between players and coalitions: an application of the Shapley Value. In A. Roth (Ed.), *The shapley value*. Cambridge: Cambridge University Press.

Bala, V., & Goyal, S. (2000). A non-cooperative model of network formation. *Econometrica, 68*, 1181–1231.

Billand, P., & Bravard, C. (2004). Non-cooperative networks in oligopolies. *International Journal of Industrial Organization, 22*, 593–609.

Bischi, G. I., & Lamantia, F. (2002). Nonlinear duopoly games with positive cost externalities due to spillover effects. *Chaos Solitons Fractals, 13*, 805–822.

Bischi, G. I., & Lamantia, F. (2004). A competition game with knowledge accumulation and spillovers. *International Game Theory Review, 6*, 323–342.

Bischi, G. I., & Lamantia, F. (2010). Knowledge accumulation in an R&D network. In T. Puu, & A. Panchuk (Eds.) *Nonlinear economic dynamics*. New York: Nova Science Publishers.

Bischi, G. I., & Lamantia, F. (2012a). A dynamic model of oligopoly with R&D externalities along networks: Part I. *Mathematics and Computers in Simulation, 84*, 51–65.

Bischi, G. I., & Lamantia, F. (2012b). A dynamic model of oligopoly with R&D externalities along networks: Part II. *Mathematics and Computers in Simulation, 84*, 66–82.

Bischi, G. I., & Naimzada, A. (1999). Global analysis of a duopoly game with bounded rationality. *Advances in Dynamic Games and Applications, 5*, 361–385.

Bischi, G. I., Dawid, H., & Kopel, M. (2003). Gaining the competitive edge using internal and external spillovers: A dynamic analysis. *Journal of Economic Dynamics and Control, 27*, 2171–2193.

Bloch, F. (1995). Endogenous structures of association in oligopolies. *The RAND Journal of Economics, 26*, 537–556.

Bloch, F. (1997). Non-cooperative models of coalition formation with spillovers. In C. Carraro, & D. Siniscalco (Eds.) *The economic theory of the environment*. Cambridge: Cambridge University Press.

Bloch, F., & Jackson, M.O. (2006). Definitions of equilibrium in network formation games. *International Journal of Game Theory, 34*(3), 305–318.

Boorman, S. (1975). A combinatorial optimization model for transmission of job information through contact networks. *Bell Journal of Economics, 6*, 216–249.

Cellini, R., & Lambertini, L. (2009). Dynamic R&D with spillovers: Competition vs cooperation. *Journal of Economic Dynamics and Control, 33*, 568–582.

Cohen, W. M., & Levinthal, D. A. (1989). Innovation and learning: The two faces of R&D. *Economic Journal, 99*, 569–596.

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly, 35*(1), 128–152.

Confessore, G., & Mancuso, P. (2002). A dynamic model of R&D competition. *Research in Economics, 56*, 365–380.

Cowan, R., & Jonard, N. (2004). Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control, 28*, 1557–1575.

D'Aspremont, C., & Jacquemin, A. (1988). Cooperative and noncooperative R&D duopoly with spillovers. *American Economic Review, 78*(5), 1133–1137.

Delapierre, M., & Mytelka, L. (1998). Blurring boundaries: new inter-firm relationships and the emergence of networked knowledge-based oligopolies. In M. G. Colombo (Ed.), *The changing boundaries of the firm*. London: Routledge Press.

Dutta, B., Van den Nouweland, A., & Tijs, S. (1995). Link formation in cooperative situations. *International Journal of Game Theory, 27*, 245–256.

Flam, S. D. (1993). Oligopolistic competition: from stability to chaos, volume 399. In F. Gori, L. Geronazzo, & M. Galeotti (Eds.), *Nonlinear dynamics in economics and social sciences*. Lecture notes in economics and mathematical systems (Vol. 399, pp. 232–237). Berlin: Springer.

Friedman, D., & Abraham, R. (2009). Bubbles and crashes: Gradient dynamics in financial markets. *Journal of Economic Dynamics and Control, 33*, 922–937.

Furth, D. (1986). Stability and instability in oligopoly. *Journal of Economic Theory, 40*, 197–228.

Gauvin, S. (1995). Networks of innovators: Evidence from canadian patents. *Group Decision and Negotiation, 4*, 411–428.

Goyal, S. (2007). *Connections: An introduction to the economics of networks*. Princeton: Princeton University Press.

Goyal, S., & Joshi, S. (2003). Networks of collaboration in oligopoly. *Games and Economic Behavior, 43*(1), 57–85.

Goyal, S., & Moraga-Gonzales, J. L. (2001). R&D networks. *The RAND Journal of Economics, 32*(4), 686–707.

Hagedoorn, J. (2002). Inter-firm R&D partnerships: An overview of major trends and patterns since 1960. *Research Policy, 31*(4), 477–492.

Hagedoorn, J., & Schakenraad, J. (1990). *Alliances and partnerships in biotechnology and information technologies*. The Netherlands: MERIT, University of Maastricht.

Jackson, M., & Wolinsky, A. (1996). A strategic model of social and economic networks. *Journal of Economic Theory, 71*, 44–74.

Kalai, E., Postlewaite, A., & Roberts, J. (1979). Barriers to trade and disadvantageous middleman: non-monotonicity of the core. *Journal of Economic Theory, 19*, 200–209.

Kamien, M. I., Muller, E., & Zang, I. (1992). Cooperative joint ventures and R&D cartels. *American Economic Review, 82*, 1293–1306.

Katz, M. (1986). An analysis of cooperative R&D. *RAND Journal of Economics, 17*, 527–543.

König, M. D., Battiston, S., Napoletano, M., & Schweitzer, F. (2011). Recombinant knowledge and the evolution of innovation networks. *Journal of Economic Behavior and Organization, 79*(3), 145–164.

König, M. D., Battiston, S., Napoletano, M., & Schweitzer, F. (2012). The efficiency and stability of R&D networks. *Games and Economic Behavior, 75*(2), 694–713.

Kopel, M., Pezzino, M., & Brand, B. (2015). Strategic location choice, R&D, and sourcing strategies. In P. Commendatore, S. Kayam, & I. Kubin I (Eds.), *Complexity & geographical economics: Topics and tools*. Berlin: Springer.

Kranton, R., & Minehart, D. (2001). A theory of buyer-seller networks. *American Economic Review, 91*, 485–508.

Leahy, D., & Neary, J. P. (1997). Public policy towards R&D in oligopolistic industries. *American Economic Review, 87*, 642–662.

Meagher, K., & Rogers, M. (2004). Network density and R&D spillovers. *Journal of Economic Behavior and Organization, 53*(2), 237–260.

Nelson, R. R. (1982). The role of knowledge in R&D efficiency. *Quarterly Journal of Economics, 97*(3), 453–470.

Petit, M. L., & Tolwinski, B. (1996). Technology sharing cartels and industrial structure. *Journal of Industrial Organization, 15*, 77–101.

Petit, M. L., & Tolwinski, B. (1999). R&D cooperation or competition? *European Economic Review, 43*(1), 185–208.

Petit, M. L., Sanna-Randaccio, F., & Tolwinski, B. (2000). Innovation and foreign investment in a dynamic oligopoly. *International Game Theory, 2*(1), 1–28.

Qiu, L. D. (1997). On the dynamic efficiency of bertrand equilibria. *Journal of Economic Theory, 75*, 213–229.

Ray, D., & Vohra, R. (1997). Equilibrium binding agreements. *Journal of Economic Theory, 73*, 30–78.

Spence, M. (1984). Cost reduction, competition and industry performance. *Econometrica, 52*, 101–121.

Suzumura, K. (1992). Cooperative and noncooperative R&D in an oligopoly with spillovers. *American Economic Review, 82*(5), 1307–1320.

Szabó, G., & Fáth, G. (2007). Evolutionary games on graphs. *Physics Reports, 446*, 97–216.

Vega-Redondo, F. (2007) *Complex social networks*. Cambridge: Cambridge University Press.

Westbrock, B. (2010). Natural concentration in industrial research collaboration. *RAND Journal of Economics, 41*(2), 351–371.

Yi, S. (1997). Stable coalition structures with externalities. *Games and Economic Behavior, 20*, 201–237.

Strategic Location Choice, R&D, and Sourcing Strategies*

Michael Kopel, Mario Pezzino, and Björn Brand

Everything is related to everything else, but closer things are more related than distant things.

Tobler (1965)

Abstract In the following chapter we figure out theoretical approaches in industrial organization to analyze firm's strategic location choice decision in an oligopolistic environment. Our research interest is twofold. First, we consider firm's investments and activities regarding R&D. Second, we focus on a firm's sourcing channel choice. In both streams of literature our main interest is put on firms' strategic interaction by means of the analysis of a firm decision's influence on its competitors' strategies. Due to the fact that this chapter does not solely display a review of the existing literature, we additionally highlight research questions and possible extensions for further research. In summary, we show the significance of firm's strategic motives within their decision-making process for both itself and its rivals. Additionally, the influence of strategic effects plays an important role for policy makers.

*We are grateful for detailed comments of an anonymous referee.

M. Kopel • B. Brand
Department of Organization and Economics of Institutions, University of Graz, Universitätsstraße 15/E4, 8010 Graz, Austria
e-mail: michael.kopel@uni-graz.at; bjoern.brand@uni-graz.at

M. Pezzino (✉)
School of Social Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, UK
e-mail: mario.pezzino@manchester.ac.uk

P. Commendatore et al. (eds.), *Complexity and Geographical Economics*,
Dynamic Modeling and Econometrics in Economics and Finance 19,
DOI 10.1007/978-3-319-12805-4_12

1 Introduction

Modern firms frequently have to make vital decisions concerning the geographic location of core activities, such as retail, production, research & development (R&D), distribution, and services. Two recent examples demonstrate that the choice of location ranks high on a management's agenda in terms of strategic value. After evaluating 17 sites in eight countries, the Austrian steel company Voestalpine AG decided to build a direct reduction plant in San Patricio County near the city of Corpus Christi in Texas (USA). CEO Wolfgang Eder is quoted as saying that this location "...was the most convincing in terms of all the key criteria, including logistics, energy supply, a well-educated workforce, and the political environment". The site is strategically located on Corpus Christi Bay and has direct access to the sea (Lundeen 2013; Voestalpine Press Release 2013). A second example which underlines the importance of strategic location choice is AstraZeneca's £330m investment to build a global headquarters in Cambridge (BBC News England 2013). Pascal Soriot, the biotech firm's CEO, explains that the new location "...will enable us to accelerate innovation by improving collaboration [...] and speeding up decision-making. The strategic centres will also allow [...] easy access to leading-edge academic and industry networks, scientific talent and valuable partnering opportunities".

The European market offers many examples of business clusters and districts, i.e. geographic areas within which groups of firms that are interconnected by horizontal and vertical relations locate to exploit some form of economic synergies and to minimize transaction costs. For example, providers of a customized input might want to be closely located to the downstream firm which would be the only purchaser of the input. At the same time, locating too close to a rival firm might prove undesirable due to a possible increase in competition (e.g. consumers could easily switch between firms). In response to trade liberalization and due to technological innovations which have lead to a dramatic reduction in transportation costs, in the 1980s and 1990s many multinational firms moved production abroad in order to exploit cost advantages. These off-shoring and outsourcing activities were oftentimes solely based on the lower cost of labor in foreign developing countries. However, more recently firms are reconsidering their previous decisions and are starting to "reshore" their activities. For example, General Electric moved manufacturing of washing machines, fridges, and heaters from China to Kentucky (USA), and Caterpillar is opening a new factory in Texas. Accounting for the hidden costs of outsourcing and off-shoring, three reasons are advanced for explaining this trend (Booth 2013). First, the labor cost advantage has been significantly reduced and the costs of transportation are rising. Instead of cheap labor, multinationals are now considering direct investments abroad to take an advantageous position in conquering fast growing foreign markets like China. Second, firms are realizing that keeping e.g. manufacturing and R&D at distant locations can have negative consequences on innovation and quality. Moreover, large distances disproportionately increase the risk of supply chain disruptions. Third and finally, companies are

starting to reshore to be able to customize their products to local markets and to be able to quickly respond to customs and preferences.

In this chapter we highlight theoretical approaches in industrial organization which shed light on a firm's strategic location choice under oligopolistic competition, where we put particular emphasis on a firm's R&D activities and a firm's choice of sourcing channel. The purpose of this chapter is not to provide a review of the contribution in the broad field of planning theory, i.e. a comprehensive analysis of the determinants and dynamics of the formation of geographic organizations. Our interest resides instead in particular aspects of location theory,[1] i.e. the analysis of the location decisions of (economic) agents such as firms, consumers, and input providers, with a particular attention to a firm's *strategic* decisions. We are using the term "strategic" intentionally. Many contributions to the literature on location theory try to identify the optimal location for an agent, given the characteristics of the agent and the features of the available geographic locations. Certainly for many disciplines, including economics and management strategy, it is very important to identify the connection between the location of economic agents and particular (exogenous) determinants, such as the technological and socioeconomic features of the possible locations. However, if the strategic interactions among agents are neglected, location decisions of economic agents can be reduced to the solution of a simple (linear) programming problem. The optimal location can be identified by optimizing an objective function that depends on the given features of the agents and the locations. A fascinating example of "non-strategic" location theory is offered by Lagrangian Points, i.e. stable locations in space that would allow space stations to restore their position after a perturbation due to an impact.[2] Economic agents however can not be compared to stationary space stations. Economic agents *react strategically* to the actions taken by other agents. For example, if a firm changes its location, it is reasonable to assume that this action would have an impact on consumers' choices and consequently on the profitability of rivals, who in turn would react by modifying their locations. Linear programming can not be of help in such a strategic environment. The toolkit necessary to analyze the strategic interaction among economic agents is offered by Game Theory. The players, i.e. economic agents like firms or consumers, take actions targeting the maximization of a payoff (or objective) function, taking into consideration that their payoff depends also on the actions taken by the other players in the game. For example, in the standard Cournot duopoly game the profits of a firm depend on the quantity decisions of *both* firms in the market. In general, the solution of a game will be given by a vector of actions for each player, where no agent has incentive to unilaterally deviate from these equilibrium strategies.[3] Game Theory is at the heart

[1] See Chan (2011).

[2] Another illustrative example for "non-strategic" location choice is the decision where to locate an international airport.

[3] For a rigorous analysis of different types of games and related equilibrium solutions, see Fudenberg and Tirole (1991).

of the economics field of Industrial Organization, i.e. the economics discipline that focuses on the analysis of imperfect markets and strategic behavior. The Industrial Organization literature offers many contributions that, since the early 1990s, have studied firms' locational strategic decision.

Before we proceed, we further restrict the focus of our interest. The present chapter will not provide an overview of the literature on strategic location choice. There are already excellent surveys on the topic, and the interested reader is referred to them. For example, the three volumes on the Economics of Location edited by Greenhut and Norman (1995a,b,c) give a broad picture of the evolution of the field and incorporates earlier and more recent papers. The recent article by Biscaia and Mota (2013) provides a historical account of location choice theory and gives a comprehensive survey of game-theoretic models of spatial competition between firms. Kilkenny and Thisse (1999) give a more selective survey of spatial economic theory, but also discuss earlier theories and general equilibrium models of location. Finally, Dembour (2008) focuses on game-theoretic models of competing regional governments, which try to influence firms' location choices by their tax/subsidy selections and infrastructure policy decisions.

The purpose of this chapter is twofold. First, we describe the contributions in the literature that study two aspects of strategic location theory that are particularly important in modern global markets, i.e. investment in R&D and innovation and the location decision of essential business activities in an international context. The location of the R&D activity should take into consideration the existence of possible positive externalities (i.e. spillovers) that the proximity to a rival firm or supplier can produce. Section 2 introduces a well-established and widely used model of strategic location decision and describes how economists have recently tried to introduce R&D spillovers into the model. Section 3 takes a more international perspective and describes some of the more recent and influential contributions in the literature studying the strategic make-or/and-buy decisions of multinational firms in a global market. The second objective of this chapter is to indicate some promising venues for future research and in particular possible extensions to the models discussed. Sections 2 and 3 both provide a discussion of possible extensions that could be studied. The final section of the chapter concludes.

2 Location Theory and Endogenous R&D Spillovers

2.1 Hotelling Line and Developments

The Industrial Organization (IO) literature offers a vast number of contributions that study firms' non-price forms of competition. Among these contributions a central role is played by the works that consider firms' decisions in terms of product

differentiation. Based on Lancaster's (1979) seminal contributions,[4] the literature distinguishes[5] between *vertical* product differentiation and *horizontal* product differentiation. Vertical differentiation considers competition among goods that unambiguously can be ranked by consumers in terms of desirable characteristics. Consequently, if a good offers higher levels of all characteristics compared to a rival product in the market, then all consumers unambiguously will prefer the former if goods were sold at the same price. *Quality* competition is a common application of the principle of vertical product differentiation.[6]

Horizontal differentiation considers instead competition among goods that can not be unambiguously ranked in terms of their characteristics and consumers may have different preferences with respect to those characteristics. For example, a manufacturer may offer a fast car with low safety standards and a rival firm may offer a slower car with higher safety standards. If cars were sold at the same price, some consumers (those who have a strong preference for speed) would buy the former while others (those with a strong preference for safety) will buy the latter. Product *variety* is a common application of the principle of horizontal product differentiation.[7] The models that focus on horizontal product differentiation, as we are going to discuss in more detail below, can also be often interpreted as models of strategic location decisions. The characteristic space can be interpreted as geographic location space. The disutility to purchase a product that does not perfectly match the preference of a consumer can be translated into the concept of transportation costs. Finally, the product variety decision of a firm can be interpreted as the location of productive activities.

A seminal (probably the most famous and most widely used) model on horizontal product differentiation is introduced in Hotelling (1929). The author studies a market represented by a linear city,[8] assumed of length 1. Consumers are assumed

[4]Lancaster's characteristics approach is based on the idea that goods are perceived and evaluated by consumers as bundles of multiple characteristics that serve the purpose of satisfying needs.

[5]The literature also distinguishes between two different approaches to product differentiation. The *non-address* approach and the *address* approach. The former considers forms of non-localized competition where each good competes simultaneously and equally with all rival varieties in the market. The latter instead considers localized competition, where each good competes with a subset of the goods in the market, e.g. those varieties that share similar characteristics. In this chapter we are going to focus exclusively on models that follow the address approach.

[6]See for example Shaked and Sutton (1982) and Motta (1993).

[7]Clearly many real markets offer products that are differentiated both vertically and horizontally. The rigorous study of contributions that consider simultaneously both forms of differentiation is out of the scope of this chapter. See for example, Economides (1989), Neven and Thisse (1990) and Irmen and Thisse (1998). The common message from these contributions is that in equilibrium firms will tend to maximize differentiation over one (dominant) dimension and minimize differentiation over other dimensions. Some of the works that we are going to discuss may be interpreted as models of variety and quality competition. We will highlight the contributions with this kind of feature.

[8]The model can be interpreted both in terms of geographic location and product variety competition. For simplicity and in line with the main purpose of this chapter we shall principally refer to the geographic location metaphor.

to be uniformly distributed over the segment and they have to incur a transportation cost[9] if the good they want to buy is offered at a shop located at a positive distance from their own location. Specifically the model considers the following assumptions:

1. A unit mass of consumers are uniformly distributed on a segment of unit length.
2. Two firms are active in the market and they sell homogeneous goods, produced at constant marginal costs, assumed for simplicity equal to zero. Firm i, $i = 1, 2$ is located at $x_i \in [0, 1]$. Without loss of generality let us assume that $x_2 \geq x_1$.
3. Consumers demand inelastically one unit of good and obtain a benefit equal to $v > 0$ from consumption. In addition, consumers incur quadratic[10] transportation costs to access a firm located at a positive distance.
4. v is sufficiently high to ensure full market coverage in equilibrium.
5. Firms target profit maximization and play a two-stage non-cooperative game. In stage one, both firms simultaneously select locations, x_i, $i = 1, 2$. In stage two, both firms charge prices, p_i, $i = 1, 2$ to consumers.

The model is solved by backward induction. The equilibrium concept is subgame perfection. The (indirect) utility of the generic consumer located at $x \in [0, 1]$ purchasing from firm i is given by:

$$U_i(x) = v - p_i - t(x - x_i)^2, \tag{1}$$

where $t(x - x_i)^2$ denotes the transportation cost and t is a transportation cost parameter.

In stage two of the game, for given firm locations, firms choose prices and consumers decide which firm to buy from comparing the full price they have to incur, $p_i + t(x - x_i)^2$. Figure 1 shows that full price incurred for any generic consumer x.

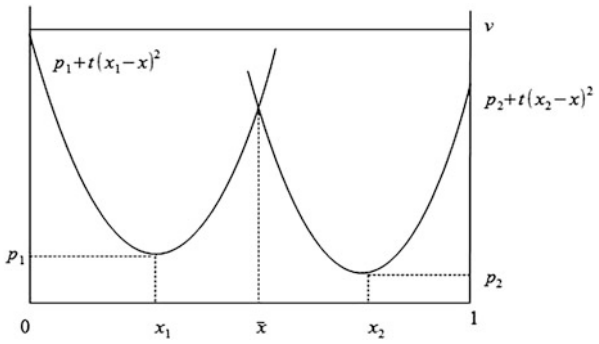
For each pair of prices chosen by the firms it is possible to identify the location of the *marginal consumer* $\bar{x}(p_1, p_2, x_1, x_2)$, i.e. the consumer indifferent to buy from firm 1 or 2. Mathematically, $\bar{x}$ is obtained as the solution of $p_1 + t(\bar{x} - x_1)^2 = p_2 + t(\bar{x} - x_2)^2$. Given the assumptions of the model, firms' demands are therefore given by

$$D_1 = \bar{x}(p_1, p_2, x_1, x_2), \qquad D_2 = 1 - D_1, \tag{2}$$

[9] According to the product variety interpretation of the Hotelling model, the transportation cost could be interpreted as the disutility that a consumer incurs when buying a product that does not perfectly match this consumers' preferences.

[10] In his original paper Hotelling considers only the case of linear transportation costs. D'Aspremont et al. (1979) however show that the model has no equilibrium (in pure strategies) under linear costs. The authors find that a unique equilibrium in pure strategy exists under the (more realistic) assumption of quadratic transport costs.

Fig. 1 Full price comparison



with $\bar{x}(p_1, p_2, x_1, x_2) = \frac{x_1+x_2}{2} + \frac{1}{2t(x_2-x_1)}(p_2 - p_1)$. For given locations, firms in stage two choose prices to maximize profits (given by $\Pi_i = p_i D_i$). It can be shown that solving the first order conditions yields the following price equilibrium in stage two,

$$p_1 = t\,(x_2 - x_1)\left(\frac{2 + x_1 + x_2}{3}\right), \qquad p_2 = t\,(x_2 - x_1)\left(\frac{4 - x_1 - x_2}{3}\right). \tag{3}$$

In stage one, firms choose locations along the segment where the decision is guided by two opposite forces. On the one hand, if firms move away from each other, (price) competition is relaxed[11] and firms can charge prices above marginal costs (here assumed equal to zero). On the other hand, locating closer to the center of the segment and therefore closer to the rival has a positive effect on the market share of each firm. D'Aspremont et al. (1979) show that with quadratic[12] transportation costs the former force prevails. Technically, $\frac{\partial \Pi_1}{\partial x_1} > 0$ and $\frac{\partial \Pi_2}{\partial x_2} < 0$, and firms in equilibrium locate at the extreme points of the segment, i.e. *maximal product differentiation* occurs.[13] Choosing different location enables firms to differentiate their products and exploit some degree of market power over the consumers more closely located to them. The degree of market power of course will depend also on the level of the transportation costs, i.e. on the parameter t. For t sufficiently close

[11]If firms are located at a strictly positive distance, then due to the existence of transportation costs their products and services are differentiated in the eyes of consumers. If firms were located at the same identical location, then their products would be homogeneous and standard Bertrand competition would bring prices down to marginal costs.

[12]Economides (1986) shows precisely how equilibrium locations depend on the convexity of the transportation cost function.

[13]The original model in Hotelling (1929) with linear transportation costs instead argues that firms would choose *minimal product differentiation*, i.e. they would both locate at the center of the segment. Such locations cannot be an equilibrium since it would create an intensely competitive environment in which firms would undercut the rival price.

to zero, competition would resemble standard Bertrand competition. If t instead is sufficiently high, the two firms become local monopolists.

The linear city model that we have just studied has some advantages and disadvantages. Among the advantages, it is a tractable duopoly model that allows researchers to study location/variety competition. The model can be easily modified to consider mixed duopoly[14] and international trade.[15] Among the disadvantages of the model, we have to mention the fact that it is essentially a duopoly story. The addition of a third firm drastically complicates the analysis and oftentimes one has to resort to numerical simulations to obtain insights on the firms' optimal strategies. It follows that the model is not helpful to study entry decisions in a market. Other limitations of the model are that product differentiation/location is assumed to be one-dimensional and that consumers are assumed to demand inelastically one unit of good.[16]

The literature offers many interesting extensions that allow economists to overcome some of the issues related to Hotelling's framework or to study different aspects of competition in products/varieties. For example, Salop (1979) considers competition between an endogenous number (possibly greater than two) of firms that locate symmetrically on a circular city.[17] Entry of a number of firms larger than two is considered also in the spokes model (see Chen and Riordan 2007). The model considers competition among firms that locate on a market represented by a number of intersecting Hotelling lines. Dos Santos Ferreira and Thisse (1996) allow Hotelling duopolists to strategically choose the transportation cost $t_i, i = 1, 2$ (for example allowing firms to reduce/increase the size and weight of their goods). The firm with the lowest transportation cost would be the vertically superior in the eyes of consumers. Finally, another interesting extension is offered by Lambertini (1997), where firms are not constrained to choose locations in the set $[0, 1]$ and set the timing of moves before competing in (unconstrained) locations and prices. The endogenous timing allows first movers to obtain a strategic advantage similar to vertical product differentiation.

[14]Mixed oligopoly is a form of imperfect competition in which firms may pursue the maximization of different objective functions. For example a private (profit maximizing) firm might compete with a public (welfare maximizing) rival. Cremer et al. (1991) study location and price competition on the linear city when one firm targets welfare maximization. They show that the presence of a public firm may be socially harmful when the number of firms in the market is greater than two and less than six.

[15]See for example Hansen and Nielsen (2006).

[16]A consequence of this assumption is that social welfare in the model essentially coincides with an inverse measure of the transportation costs incurred by all consumers.

[17]Bouckaert (2000) and Madden and Pezzino (2011) extend the Salop model allowing a firm to take a location in the center. This extension allows for the study of markets where high street shops compete with Internet/mail order rivals.

2.2 Endogenous R&D Spillovers

In this subsection we focus on a recent extension of the Hotelling model in which firms through their location decisions can affect the level of R&D spillovers. The vast economic literature on R&D distinguishes between different forms of innovation. To keep it to a minimum, *process innovation* considers the ability of a firm to invest in R&D in order to find more efficient ways to produce a product. A standard output of process innovation is a cost-reducing invention. *Product innovation* instead considers the ability of a firm to invest in R&D in order to change the characteristics of a product or to produce a completely new version of it.[18] The contributions of this literature consider many different aspects of innovation. Among them, the role played by patents and licenses, the ownership of the rights over the inventions, and the risk and timing of research. A very peculiar and interesting feature of R&D is that often the research undertaken by a firm might have positive externalities on the research (in terms of lower costs or higher probability of positive outcome) of the rivals. In other words in many industries *research spillovers* might play a very important role. See also the chapter on R&D networks by Bischi and Lamantia (2015) in this volume.

In the literature, there are two main approaches[19] to the study of R&D spillovers. Following Kamien et al. (1992) spillovers have an effect on R&D *inputs*. In other words, in the presence of spillovers the effective investment in R&D of a firm depends on the firm's own expenditure plus a fraction of the rivals' expenditure. The other approach to modeling spillovers is offered by d'Aspremont & Jacquemin (1988) where the outcome of research efforts of each firm depends also on the *output* of the other firms in the market. Both papers are interested in the role played by spillovers in the formation and performance of R&D collaboration among firms. The papers study a symmetric two-stage game where firms in stage one select R&D strategies and in stage two offer homogeneous goods and compete in quantities. The common message is that with sufficiently large spillovers, R&D collaboration (for example via an R&D cartel) produces more innovation and results in lower prices. The basic intuition is that cartel members internalize the research externality on aggregate profits.

In what follows we focus on the approach suggested by Kamien et al. (1992). The output of R&D for a generic firm i that competes in the final market with firm j is given by:

$$R_i = r_i + \beta r_j, \tag{4}$$

[18] Additionally, a firm's investment in R&D strengthens its ability to identify and assimilate information from the market to imitate and exploit already existing knowledge from others. This further innovation and learning incentive is labeled *absorptive capacity*, see Cohen and Levinthal (1989).

[19] See Amir (2000) and Amir et al. (2008) for a critical comparison of the models.

where r_z are the R&D investments of firm $z \in \{1, 2\}$. The exogenous parameter $\beta \in [0, 1]$ defines the extent of R&D spillovers. Contributions that follow this approach are often interested in studying how firms' behavior (and consequently the applicability of different economic policies) is affected by β.

While most of the contributions in the literature focus on the study of horizontal R&D spillovers, i.e. research externalities among rivals, few papers consider the spillovers created by vertical relations. If both, the suppliers of specialized inputs and the producers of final products that use these inputs, have to undertake R&D projects, then vertical spillovers (i.e. research externalities between suppliers and manufacturers) may take place. Similar to the case of horizontal externalities, different forms of cooperation and coordination can be considered by firms in order to control or exploit the vertical spillovers. Ishii (2004), for example,[20] studies competition in a market where two producers offer homogeneous goods and compete in quantities for final consumers. Both firms experience horizontal R&D spillovers and require the provision of an input. It is assumed that both firms acquire the input in equal quantities from two competing suppliers. Like the downstream firms, the suppliers of the input invest in (cost-reducing) R&D. All firms in the market experience both horizontal and vertical spillovers in R&D. In other words, the ability of a firm to reduce its costs depends on its own investments, on the effort of the rival (horizontal spillovers) and the effort of the suppliers that belong to the vertical relationship. The author considers the following three-stage game. In stage one firms and suppliers choose R&D investments. In stage two the input suppliers compete in quantities. In stage three the manufacturers compete in quantities in the final product market. The author shows that *irrespective of the extent of the spillovers* vertical coordination produces higher innovation and is always socially desirable compared to non-cooperative R&D. The result is striking if compared with the common message of D'Aspremont and Jacquemin (1988) and Kamien et al. (1992) where vertical spillovers are absent: here it is argued that collaboration boosts innovation only when firms experience sufficiently large spillovers. The intuition for the difference in results is that due to vertical spillovers the investment in R&D of a manufacturer increases the demand for the input and reduces the costs of the supplier and the same reasoning can be applied for the investment of the supplier. Collaboration or integration[21] allow the firms to internalize these effects and consequently achieve a level of cost reduction greater than under competition.

Since spillovers may play such a key role in shaping competition for the final consumers, it is natural to ask whether firms may find any way to control and influence the extent of the effects that own research might have on the rival's

[20] See also Atallah (2002).

[21] The author considers three forms of coordination: Vertical R&D cartels, vertical non-cooperative Research Joint Ventures, RJVs, and vertical RJV cartels. In the first case firms coordinate R&D decisions. In the second case firms share R&D knowledge (i.e. maximize the vertical spillovers). In the last form of cooperation, firms both share knowledge and coordinate their R&D investments.

ability to produce innovation. The literature offers various attempts[22] to deal with endogenous spillovers, in particular considering the possibility that firms' R&D activities could affect the parameter β.

Piga and Poyago-Theotoky (2005), PPT from now on, offer a very interesting and tractable extension of the Hotelling (1929) model with quadratic transportation costs that allow to study endogenous R&D spillovers. The authors argue that in many real markets the extent of spillovers is directly related to the proximity of firms.[23] This way the authors introduce a strategic force that works against the principle of maximal product differentiation. While it would be still true that firms have an incentive to locate at a positive distance in order to relax price competition, now such a distance may not be maximized in equilibrium since firms will try to exploit the positive effects of R&D spillovers. Specifically PPT consider the following utility function for the generic consumer located at $x \in [0, 1]$ who buys from firm i:

$$U_i(x) = v + R_i - p_i - t(x - x_i)^2. \tag{5}$$

Consumer utility increases with the output of R&D (we can interpret R_i as the perceived quality of product i). As we described in the previous section, the possibility that firms may invest in quality while providing horizontally differentiated goods is not novel. The novelty resides in the following assumption. PPT assume that R_i takes the following functional form:

$$R_i = r_i + (1 - x_2 + x_1)r_j. \tag{6}$$

Expression (6) says that if firms share the same location, i.e. $x_1 = x_2$ then the spillovers are maximized. The quality perceived by consumers is the same for the two goods and given by the sum of the firms' investments in R&D. If firms instead would maximize their distance, i.e. $x_1 = 0$ and $x_2 = 1$, then spillovers would disappear. The case without spillovers has been already studied in the literature (see for example Calem and Rizzo 1995; Brekke et al. 2006; Barros and Martinez-Giralt 2002). The model would be essentially an extension of the standard Hotelling framework with the addition of a quality selection stage in the game. Firms select both qualities and locations/varieties, under the assumption that consumers share the same willingness to pay for quality. This assumption makes the model tractable compared to the standard vertical differentiation framework (as in Motta 1993, where the equilibrium produces asymmetric qualities) and ensures the existence of a symmetric equilibrium.

[22] See for example De Fraja (1993), Amir et al. (2003) and Milliou (2009). An interesting form of affecting the degree of R&D spillovers is also created with the use of research joint ventures. When the joint venture requires the use of a single lab then spillovers are maximized.

[23] In terms of product space, it would be reasonable to suppose that spillovers may be facilitated if firms produced very similar varieties.

The R&D related costs are assumed to be equal to $c(r_i) = \frac{1}{2}r_i^2$. The timing of the three-stage game that the authors consider is as follows. In stage one firms choose locations, $x_i \in [0, 1]$, $i = 1, 2$. In stage two firms choose R&D investments, $r_i \in [0, \infty)$, $i = 1, 2$. In the final stage of the game, firms compete in prices. The authors focus on symmetric locations, i.e. $x_1 + x_2 = 1$, and the game is solved by backward induction.

In equilibrium[24] the two firms choose extreme locations for high levels of t. For intermediate levels of t instead the two firms choose interior (and symmetric) locations. Moreover, equilibrium R&D investments are symmetric. In equilibrium, therefore firms experience minimal quality differentiation and strictly positive (maximal for large t) horizontal differentiation.[25]

The model in PPT offers a tractable framework for the study of strategic location choices and endogenous R&D spillovers. In the following subsection we consider a number of directions for further research.

2.3 Extensions

We have discussed so far the importance of strategic location decisions, stressing in particular the role that proximity between competing firms may play in the provision of innovation. Specifically we have identified in Piga and Poyago-Theotoky (2005) an interesting and tractable way to introduce endogenous R&D spillovers related to firms' locations. In this subsection we discuss some paths for future research that may prove fruitful and interesting.

2.3.1 Strategic Delegation

Vickers (1985) and Fershtman and Judd (1987) are the first and seminal papers on the use of *strategic delegation*. The idea is that in an imperfectly competitive environment, firms may decide to hire managers in order to commit to a more or less aggressive competitive behavior in the market. In the standard Cournot model where quantity decisions are delegated to managers who are compensated based on profits *and* sales revenues, in effect both firms try to obtain a Stackelberg leader position and sell a higher quantity compared to the standard Cournot-Nash equilibrium without delegation. Zhao (2012) argues that delegation can be implemented also based on other strategic variables, such as R&D investments (see also Kopel and Riegler 2006). The author considers a four-stage duopoly game played by two

[24]It is assumed that the condition $t > 2/9$ is satisfied. This assumption ensures that stability and second order conditions are met in equilibrium.

[25]The result is in line with contributions that study bi-dimensional product differentiation. Firms in general maximally differentiate in one dimension and minimally differentiate in the other.

identical firms. In stage one, the *owner* of each firm simultaneously chooses the location of the firm on a Hotelling line. In stage two, without delegation the owners invest in quality-enhancing R&D and with delegation this decision is taken by managers. If in stage two there is no delegation, then in stage three the owners decide whether to delegate price decisions to managers. In stage four without delegation owners compete in prices and with delegation managers choose prices. The author defines *semi-delegation*, a situation in which managers take only price decisions, and *full delegation*, the case in which owners only choose locations.

Managers are assumed to be concerned only about their financial position and are offered an incentive contract that targets both profits and market shares. In addition, R&D spillovers identical to those described in PPT are considered. It follows that if no firm decides to delegate, then the model coincides with the analysis in PPT. The author shows that under semi-delegation managers are less aggressive (meaning higher prices and profits). Firms tend to locate further away from the rival and invest more heavily in R&D. Full delegation, however, is more socially efficient. Finally it is shown that semi-delegation is a Pareto dominant strategy for both firms. The model could be extended in various interesting directions. Different contractual forms for the managers could be considered (e.g. performance measures based on output, sales revenue, or relative profits). The case of full delegation could also include the possibility that location decisions are taken by managers as well. Another extension could be to study how firms behavior changes once one of them is a public firm[26] or socially concerned firm (e.g. Kopel and Brand 2012). The public firm would target some measure of welfare and/or introduce it in the performance measure of the manager's incentive contract. In the next subsection we explore mixed competition in more detail.

2.3.2 Mixed Oligopoly

In many markets firms with different mission and different objective functions compete for the same set of consumers. The literature on mixed oligopolies has studied how location decisions (both on the Hotelling line and the Salop circle[27]) are affected by the presence of a public firm. It is in general assumed that the public (or socially responsible) firm targets the maximization of an objective function other than pure profits. It is assumed that public firms maximize social welfare, consumer surplus or a mix of own profits and the surplus of other economic agents in the market. The symmetric nature of the model presented in PPT could be altered introducing the possibility that one of the two firms targets the maximization of a composite function such as:

$$G_i = \Pi_i + \alpha CS_i, \tag{7}$$

[26] See Sect. 2.3.2.

[27] See Matsumura and Matsushima (2003, 2004) and Matsushima and Matsumura (2003).

where Π_i represents the profits of firm i, CS_i is the surplus of consumers served by i and $\alpha \in [0, 1]$ is a parameter that indicates the degree of social concern of the firm. It would be interesting to study how firms' decisions, in particular in terms of location and investments in R&D are affected by the public nature of one of the two rivals.[28]

2.3.3 Salop Circular City

Salop's circular city model allows to study entry decisions in an oligopoly model of horizontal product differentiation (Salop 1979). It would be interesting to introduce the concept of endogenous spillovers as described in PPT in the circular city framework. Doing so it would be possible to study how the presence of endogenous spillovers affect entry decisions. In particular, if spillovers were increasing with the number of firms in the market, then entry may have two socially desirable effects. It would decrease the aggregate measure of the transportation costs (as in the standard model developed by Salop 1979) and increase the aggregate level of perceived innovation/quality for given quality investments. If these positive effects more than offset the negative *business stealing* effect of entry, the Salop result of excessive entry may be reversed.[29]

In a circular city model, R&D spillovers may take place in different forms. The spillovers may be simply increasing in the number of firms in the market (and consequently in the proximity of firms). For example, the effective quality perceived by consumers buying from firm i might be given by

$$Q_i = q_i + (1 - \frac{1}{n})\bar{q}_{-i}, \tag{8}$$

where n is the number of firms that entered the market, q_i is the quality investment by firm i, and $\bar{q}_{-i}$ is some measure of the aggregate quality provided by the other $(n - 1)$ firms. Depending of the particular market considered, $\bar{q}_{-i}$ could be the average quality or the summation of all the qualities.

The presence of central firm, as described in Madden and Pezzino (2011) may be another source of endogenous spillovers. See Fig. 2 for a graphical intuition. The central firm is a *global* competitor, in the sense that it competes with all the

[28] A recent paper by Zhang and Li (2013) considers the case with $\alpha = 1$ and a timing in which firms choose locations on a line *after* having invested in R&D. The paper shows that the R&D activity of the public firm is strongly affected by the degree of spillovers and privatization may be socially undesirable.

[29] Mankiw and Whinston (1986) introduced the concept of business stealing. Due to the inelastic demand and full market coverage assumptions, when a firm enters the market it incurs a fixed cost and *steals* some market share from neighboring firms. The private incentive to enter the market is greater than the social desirability of entry producing the excessive entry result, which occurs under Salop competition.

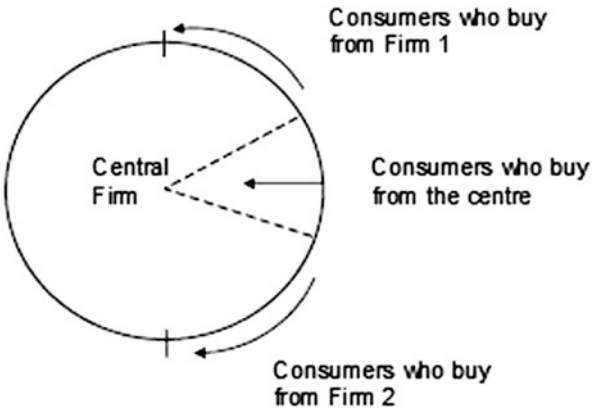


Fig. 2 Salop circle with center

other firms that are located on the perimeter of the circle. Without the central firm, competition is localized (i.e. each firm directly competes only with the two closest rivals on the circle) and spillovers may not be generated. The central firm may instead play the role of a "R&D hub". Through competition with the center, perimeter firms may benefit from the R&D investments of *all* rivals in the market. It follows that a regulator should assess the presence of a central, homogeneous product firm, not only on the basis of the effects produced on the aggregate transportation costs incurred by consumers (as considered in Madden and Pezzino 2011), but also on the quality-enhancing effects that may be created due to R&D spillovers.

3 Strategic Sourcing and Organizational Governance

The increasing importance of international outsourcing and procurement of intermediate inputs within the firm through foreign direct investments are certainly among the most controversially discussed issues in the business press and among policy makers. Trade liberalization has led to considerable growth in the vertical fragmentation of production and has led firms to outsource several core and non-core activities of their value chain. Some industry observers raised the concern about whether outsourcing makes good business sense or if outsourcing has simply gone too far (e.g. Engardio and Einhorn 2005; Uchitelle 2006). As outlined in the introduction, many multinational firms reconsider their sourcing strategies, which ultimately might lead to a reversion of the outsourcing trend.

Economic research has increasingly focused on the combination of international competition and global trade and the choice of organizational form and has tried to shed light on the drivers of foreign direct investment (FDI), outsourcing, and vertical integration. The majority of contributions along this line of research has been set in a (general equilibrium) monopolistic competition framework, but enriched

by elements of contract theory á la Grossman-Hart-Moore and transaction cost economics á la Williamson (e.g. Spencer 2005; Antràs and Helpman 2004; Du et al. 2009; Grossman and Helpman 2004, 2005). However, these articles neglect an important point which is present in many real-world industries and markets like automotive, airframe manufacturing, PC software and hardware industries: the strategic interaction among key market players. The view that multinational corporations possess market power and that market structure is dominantly oligopolistic is corroborated by Spencer (2005): "Another promising direction for research is to recognize that many firms involved in contractual outsourcing, such as IBM and General Electrics, are extremely large and have some market power. This suggests a need to understand the *strategic motives of oligopolistic firms* that engage in international contractual outsourcing" (pp. 1132–1133, emphasis added). In fields like industrial organization, economics, accounting, and production/operations management, many contributions are using game-theoretic models to study the benefits and costs of outsourcing vs. insourcing, strategic channel coordination, and its influence on the endogenous choice of organizational governance. Surprisingly, research on multinational enterprises in international competition has started to join this trend only rather recently. In this subsection we are presenting a selection of papers which deal with the sourcing strategies of multinational firms engaged in international oligopolistic competition. An interesting twist of global sourcing settings in relation to existing models in IO and production and operations management is that multinational firms have a richer array of options due to the location of the sourcing channel: a firm can produce the necessary inputs in-house, it can outsource the inputs in the home country, it can produce them in a wholly owned subsidiary that is located in the foreign country, or it can import them from a foreign supplier. Furthermore, it can combine these sourcing strategies, for example it can bi-source the same input (Du et al. 2009). This multiplicity of modes of organization makes the study of the trade-offs between them rather complex. The two branches of literature, general equilibrium approaches based on monopolistic competition where firms are atomistic and have no market power and the game-theoretic models using a partial equilibrium approach seem to be completely unrelated. Recent research shows, however, that these two views can be consolidated in a "General Oligopolistic Equilibrium (GOLE)" framework (e.g. Neary 2009; Bastos and Straume 2012). This more general framework assumes that firms are "large in the small but small in the large" (p. 4), i.e. they are significant players in their own market, interacting strategically with their local rivals, but infinitesimal in the economy as a whole (Neary 2010).

The literature on FDI has focused on two motives of firms choosing a country for their newly developed plant. The first reason is that oftentimes firms wish to have easier access to the localized markets of this country. To manufacture locally saves transportation costs and also circumvents trade barriers. A second reason is that firms want to gain access to (human and other) resources of the foreign country and use the new plant as a low-cost production facility. Likewise, outsourcing to foreign suppliers saves labor costs. The more recent literature identifies a third, and more subtle, mechanism which provides an incentive for firms to do FDI

or outsource: a strategic motive. In oligopolistic markets the action of one firm influences the payoffs of all the other firms. Consequently, FDI of a firm (foreign production instead of export) or outsourcing to a common supplier has an impact on multinational rivals and changes the profits of all firms in the foreign and home country. A proper analysis of the incentives for multinational enterprises has to take this into account.

A good example for such an analysis using a game-theoretic model is Leahy and Pavelin (2008). They consider two firms which serve the home market and can serve the foreign market either by export or, following FDI, from a local plant in the foreign country. Firms engage in corporate-wide cost-reducing R&D investment, which lowers production costs in all plants of the firm. They identify the following strategic incentive for FDI. If the firm opens a plant in the foreign country it saves on transportation costs (compared to export) and hence offers a larger quantity in the foreign market. This increases the firm's total quantity and, therefore, increases its incentives for cost-reducing R&D. In contrast, the rival firm's total output decreases and so does the incentive for cost-reducing investments. This, however, also has an effect on the market shares in the home market. The focal firm's profit in the home market is improved. Taken together, they obtain the surprising result that FDI studied in isolation (due to its fixed costs of entering the foreign market) might be unprofitable, but the improvement in profits on the home market can overcompensate these losses, which makes (strategic) FDI the preferred choice.

Chen et al. (2004) consider *strategic* outsourcing, which means that beside the usual cost-saving argument, outsourcing can also be based on a strategic motive and can be beneficial even if outsourcing is more costly than making the required input in-house (see also Kopel et al. 2013). They study a price-setting model with two firms, one domestic and one foreign, which compete in the final product market. Production of the final product requires an intermediate input. The domestic firm can either make the intermediate input or buy the input from the foreign firm. The foreign firm is vertically integrated and makes the input. Chen et al. (2004) show that the domestic firm might buy the input from the foreign firm even if the total purchasing costs (including tariffs) are higher. The reason is that buying from the rival firm induces weaker behavior of the competitor on the final product market since aggressive pricing of the rival results in lower market share of the domestic firm and hurts the rival's sales of the input.

It is common in the literature to view outsourcing (buy) and in-house production (make) as two exclusive choices. Many firms, however, practice bi-sourcing (or concurrent sourcing), meaning that they make *and* buy. Oladi et al. (2007) study strategic international bi-sourcing in a duopoly context where final goods are sold in a global competitive market. Manufacturing the final product requires inputs which are supplied to firms in the local market. Furthermore, production requires a specific input which can be bought locally, but also (additionally) procured from the rival market. The authors show that although the price of the specific input is the same in both markets, both firms employ a bi-sourcing strategy, i.e. they make and buy this input. The (strategic) reason for bi-sourcing in their model is that the presence of a firm in both input markets serves as a commitment device and influences the input prices and the rival's behavior and payoff.

3.1 FDI, Outsourcing and Hold-up

In this subsection, we turn to an interesting contribution by Leahy and Montagna (2009) and describe it in a more comprehensive way since it captures many important details of a firm's situation when it has to select between FDI or outsourcing. It also includes in a simple way elements of strategic interaction between firms in the market and the relationship between a firm and its supplier in the presence of contractual incompleteness, non-verifiable investments, and hold-up threats. Two ex ante identical firms are located in the North and sell a homogeneous final product on the North market. Inverse demand in this market is $p = a - (y_1 + y_2)$, where y_i denotes the quantity produced by firm i. Manufacturing the final product requires a specific input which either is procured from a specialized supplier i in the South (outsourcing option) or manufactured by the firm in a newly established fully owned plant in the South (FDI option). If the firm outsources, the price of the input is q_i. In case of FDI, the firm has to pay a fixed (setup) cost of F and marginal cost per unit is r. In either case, transportation costs per unit of input is t. To manufacture the final product, the specific input is combined with a composite input (whose price is 1) and the quantity of composite input needed depends on the usefulness z_i of the specific input (to be defined below). The term $e_i = \overline{e} - z_i > 0$ denotes firm i's requirement of composite input, where $\overline{e}$ is a constant. Consequently, a higher usefulness of the specific input results in a smaller quantity of composite input needed in production of the final product. The usefulness of the specific input to the final product depends on an investment, K, in quality and customization, where $z = \sqrt{K}$. Marginal production costs of firm i under the outsourcing (O) option and the FDI (I) option are given by

$$c_i^O = q_i + \overline{e} - z_i + t, \qquad c_i^I = r + \overline{e} - z_i + t, \tag{9}$$

respectively. If firm i does FDI, then profits are

$$\pi_i^I = (p - c_i^I) y_i - K_i - F. \tag{10}$$

If firm i outsources, then profits are simply

$$\pi_i^O = (p - c_i^O) y_i, \tag{11}$$

and the investment costs in quality and customization are now borne by the supplier, which has profits of

$$\mu_i = (q_i - r^m) q_i - K_i - E, \tag{12}$$

where E denote entry costs. The timing of the Leahy-Montagna-game is as follows. In the first stage, firms decide whether to outsource or to do FDI (vertical integration). In stage two, either the firm (FDI option) or the supplier (outsourcing option) chooses K_i. In stage three, in case of outsourcing the firm bargains with

their supplier over price q_i. This captures the realistic scenario that contractual incompleteness governs the relationship between the firm and its supplier and that the supplier's level of investment is not verifiable and cannot be part of any contract. In stage four, the input is shipped to the North at transport cost t and the final product is manufactured and offered in the market.

Using this rich and interesting model, Leahy and Montagna (2009) demonstrate that, again, the underlying motive for outsourcing might be strategic: it softens the behavior of the rival. Interestingly, they also show that asymmetric equilibria might arise although both firms are ex ante symmetric (on this point, see also Kopel and Löffler 2012). They argue that this explains the heterogeneity observed among multinational enterprises in international competition. Leahy and Montagna (2012) employ an identical setting to address the question if international outsourcing occurs to weaken strong domestic unions.

3.2 Extensions

As mentioned above, the literature on international oligopolistic competition has only recently employed game-theoretic models to address important issues like strategic sourcing and the optimal choice of governance of the sourcing channel. Much could be learned from the modeling approaches in IO, economics, and production and operations management. These setups could be applied to particular questions of international oligopolistic competition taking into account the wider array of governance modes of multinational firms. For example, many firms bi-source their inputs and it would be interesting to understand under which circumstances bi-sourcing is the optimal sourcing strategy. Further topics of interest include strategic licensing of R&D, sourcing from rival firms, and the optimal transfer pricing method in international competition. An interesting recent contribution is provided by Wu and Zhang (2013), who study sourcing strategies and the trade-off of backshoring previously outsourced activities. Research along these lines would provide a better understanding of the issues which multinational firms face and would enable policy makers to better address the needs of these companies who are engaged in international competition in oligopolistic markets.

4 Conclusion

Modern national and multinational companies operate in a dynamic and complex environment where firm performance is in general affected by the interplay of various factors, both structural and strategic. Environmental factors, such as the regulatory framework adopted, the type of local demand, the quality and cost of local inputs, the degree of competition in the local market, etc., own strategic decisions and (re)actions taken by current and potential business rivals all should be

simultaneously and carefully pondered by firms choosing location. The definition of the optimal geographic location of business core activities therefore entails strategic considerations that should be understood, evaluated and classified resorting to Game Theory tools. The issue is particularly important nowadays. Changes in the economic and technological structure of countries that used to be obvious business destinations for multinational firms in the 1980s and 1990s are recently creating an inversion in the trend regarding the location of essential core activities, inducing experts and academics to talk about *re-shoring* of business to the countries of origin. At least two reasons can be identified behind this reversed trend. First, domestic consumers are mature, selective and informed and require often highly specialized and personalized products and services. Second, the cost savings created by producing abroad or providing customer support located far away in other countries are shrinking and sometimes do not motivate the decision to outsource/offshore.

This chapter provided a brief and intuitive introduction of issues related to strategic location decisions, with a particular emphasis on activities such as R&D and outsourcing. The location decision of an R&D center is particularly interesting and complex, given the possible existence of R&D spillovers, i.e. the possibility that the outcome of own research may depend also on the effort of rivals located in the same geographic market. The purpose of the chapter was twofold. First, it provided a critical description of some of the most recent and influential contributions of the theoretical economic literature on strategic location, outsourcing and R&D investments. Second, after identifying economic questions still unanswered, it offered a discussion of possible promising paths for future research and analysis. Section 2 discussed the way strategic location has been considered by the mainstream Industrial Organization literature. Firms may want to locate at a positive distance from their competitors in the attempt to control a niche of the market and make the most of their market power. Distance (that in the product characteristic space would instead be considered as product differentiation) may relax competition and increase firms' profits. If R&D spillovers were favored by proximity, firms may decide not to maximize the distance between them. Section 3 took a more international perspective and considered location strategies in terms of strategic outsourcing, FDI and *make* decisions. The decision to outsource, for example from rival (foreign) firm, should not be based in general only on cost reduction, but also on the consideration of the effects that the decision might have on the strategic reaction of the rivals. FDI may be unprofitable in the short run, but it may prove to be an excellent way to enter a foreign market and acquire a valuable position in it. The chapter discussed possible extensions to the key contributions in the literature. In particular it would be interesting to study how the effects of R&D spillovers change depending on market characteristics (such as the number of firms and the structure of the possible network) and the type of firms involved (for example under mixed competition between private and public firms).[30] R&D spillovers may play a role

[30]For a detailed study of R&D networks, we refer the reader to the chapter by Bischi and Lamantia (2015) in this volume.

also in an international context, where a firm might decide to locate in a particular country not to exploit some form of cost saving, but to follow a competitor and favor the existence of R&D spillovers. The study of strategic use of *bi-sourcing* (i.e. the concurrent use of outsourcing and make in-house) would be another interesting venue for further research. In terms of policy implications and related issues, the chapter provides a key message. In order to study and assess the effect that firms' location decisions have on the economy and how different industrial and fiscal policies affect such decisions, it is essential to have a clear understanding of the strategic motives behind firms' behaviors. If in particular innovation and R&D are involved in the location decision of firms, product quality and technical/productive efficiency considerations should be carefully assessed. The issue may be further complicated by possible forms of government competition such as taxation or the definition of quality/safety standards. Favorable tax rates and infrastructures, possibly accompanied by not binding quality standards may influence the location decision of firms. When deciding upon tax rates and standards, and the possible effects of harmonization or mutual recognition, it is vital that governments and intergovernmental bodies fully understand the effects their decisions will have on location, R&D, quality, competition and, in the end, the welfare of consumers and firms.

References

Amir, R. (2000). Modelling imperfectly appropriable R&D via spillovers. *International Journal of Industrial Organization, 18*(7), 1013–1032.

Amir, R., Evstigneev, I., & Wooders, J. (2003). Noncooperative versus cooperative R&D with endogenous spillover rates. *Games and Economic Behavior, 42*(2), 183–207.

Amir, R., Jin, J. Y., & Troege, M. (2008). On additive spillovers and returns to scale in R&D. *International Journal of Industrial Organization, 26*(3), 695–703.

Antràs, P., & Helpman, E. (2004). Global sourcing. *Journal of Political Economy, 112*(3), 552–580.

Atallah, G. (2002). Vertical R&D spillovers, cooperation, market structure, and innovation. *Economics of Innovation and New Technology, 11*(3), 179–209.

Barros, P. P., & Martinez-Giralt, X. (2002). Public and private provision of health care. *Journal of Economics & Management Strategy, 11*(1), 109–133.

Bastos, P., & Straume, O. R. (2012). Globalization, product differentiation, and wage inequality. *Canadian Journal of Economics, 45*(3),857–878.

BBC News England. (2013). *Astrazeneca axes 700 jobs in Cambridge move*. http://www.bbc.co.uk/news/uk-england-21833207. Accessed 18 Mar 2013.

Biscaia, R., & Mota, I. (2013). Models of spatial competition: A critical review. *Papers in Regional Science*. doi:10.1111/j.1435-5957.2012.00441.x.

Bischi, G. I., & Lamantia, F. (2015). R&D networks. In P. Commendatore, S. Kayam, & I. Kubin (Eds.), *Complexity and geographical economics: Topics and tools*. Heidelberg: Springer (this volume). doi:10.1007/978-3-319-12805-4_2.

Booth, T. (2013). Here, there and everywhere. *Economist*. http://www.economist.com/news/special-report/21569572-after-decades-sending-work-across-world-companies-are-rethinking-their-offshoring. Accessed 8 Apr 2013.

Bouckaert, J. (2000). Monopolistic competition with a mail order business. *Economics Letters, 66* (3), 303–310.
Brekke, K. R., Nuscheler, R., Straume, O. R. (2006). Quality and location choices under price regulation. *Journal of Economics & Management Strategy, 15*(1), 207–227.
Calem, P. S., & Rizzo, J. A. (1995). Competition and specialization in the hospital industry: An application of Hotelling's location model. *Southern Economic Journal, 61*(4), 1182–1198.
Chan, Y. (2011). *Location theory and decision analysis: Analytics of spatial information technology* (2nd ed.). Heidelberg: Springer.
Chen, Y., & Riordan, M. H. (2007). Price and variety in the spokes model. *Economic Journal, 117* (522), 897–921.
Chen, Y., Ishikawa, J., & Yu, Z. (2004). Trade liberalization and strategic outsourcing. *Journal of International Economics, 63*(2), 419–436.
Cohen, W. M., & Levinthal, D. A. (1989). Innovation and learning: The two faces of R&D. *Economic Journal, 99*(397), 569–596.
Cremer, H., Marchand, M., & Thisse, J.-F. (1991). Mixed oligopoly with differentiated products. *International Journal of Industrial Organization, 9*(1), 43–53.
D'Aspremont, C., & Jacquemin, A. (1988). Cooperative and noncooperative R&D in duopoly with spillovers. *American Economic Review, 78*(5), 1133–1137.
D'Aspremont, C., Gabszewicz, J. J., & Thisse, J.-F. (1979). On Hotelling's "stability in competition". *Econometrica, 47*(5), 1145–1150.
De Fraja, G. (1993). Strategic spillovers in patent races. *International Journal of Industrial Organization, 11*(1), 139–146.
Dembour, C. (2008). Competition for business location: A survey. *Journal of Industry, Competition and Trade, 8*(2), 89–111.
Dos Santos Ferreira, R., & Thisse, J.-F. (1996). Horizontal and vertical differentiation: The Launhardt model. *International Journal of Industrial Organization, 14*(4), 485–506.
Du, J., Lu, Y., & Tao, Z. (2009). Bi-sourcing in the global economy. *Journal of International Economics, 77*(2), 215–222.
Economides, N. (1986). Minimal and maximal product differentiation in Hotelling's duopoly. *Economics Letters, 21*(1), 67–71.
Economides, N. (1989). Quality variations and maximal variety differentiation. *Regional Science and Urban Economics, 19*(1), 21–29.
Engardio, P., & Einhorn, B. (2005). Outsourcing innovation. *Business Week*. http://www.businessweek.com/stories/2005-03-20/outsourcing-innovation. Accessed 21 Mar 2013.
Fershtman, C., & Judd, K. (1987). Equilibrium incentives in oligopoly. *American Economic Review, 77*(5), 927–940.
Fudenberg, D., & Tirole, J. (1991). *Game theory*. Cambridge: MIT Press.
Greenhut, M. L., & Norman, G. (1995a). *The economics of location - volume 1: Location theory*. Cheltenham: Edward Elgar Publishing.
Greenhut, M. L., & Norman, G. (1995b). *The economics of location - volume 2: Space and value*. Cheltenham: Edward Elgar Publishing.
Greenhut, M. L., & Norman, G. (1995c). *The economics of location - volume 3: Spatial microeconomics*. Cheltenham: Edward Elgar Publishing.
Grossman, G.M., & Helpman, E. (2004). Managerial incentives and the international organization of production. *Journal of International Economics, 63*(2), 237–262.
Grossman, G. M., & Helpman, E. (2005). Outsourcing in a global economy. *Review of Economic Studies, 72*(1), 135–159.
Hansen, J. D., & Nielsen, J. U. (2006). Economic integration and quality standards in a duopoly model with horizontal and vertical product differentiation. *Journal of Economic Integration, 21*(4), 837–860.
Hotelling, H. (1929). The stability of competition. *Economic Journal, 39*(153), 41–57 (1929)
Irmen, A., & Thisse, J.-F. (1998). Competition in multi-characteristics spaces: Hotelling was almost right. *Journal of Economic Theory, 78*(1), 76–102.

Ishii, A. (2004). Cooperative R&D between vertically related firms with spillovers. *International Journal of Industrial Organization, 22*(8–9), 1213–1235.
Kamien, M. I., Muller, E., & Zang, I. (1992). Research joint ventures and R&D cartels. *American Economic Review, 82*(5), 1293–1306.
Kilkenny, M., & Thisse, J.-F. (1999). Economics of location: A selective survey. *Computers & Operations Research, 26*(14), 1369–1394.
Kopel, M., & Brand, B. (2012). Socially responsible firms and endogenous choice of strategic incentives. *Economic Modelling, 29*(3), 982–989.
Kopel, M., & Löffler, C. (2012). Organizational governance, leadership, and the influence of competition. *Journal of Institutional and Theoretical Economics, 168*(3), 362–392.
Kopel, M., & Riegler, C. (2006). R&D in a strategic delegation game revisited: A note. *Managerial and Decision Economics, 27*(7), 605–6012.
Kopel, M., Löffler, C., & Pfeiffer, T. (2013). *Sourcing strategies of a multi-product firm when sourcing multiple inputs*. Working Paper.
Lambertini, L. (1997). Unicity of the equilibrium in the unconstrained Hotelling model. *Regional Science and Urban Economics, 27*(6), 785–798.
Lancaster, K. (1979). *Variety, equity, and efficiency: Product variety in an industrial society (studies in economics)*. New York: Columbia University Press.
Leahy, D., & Montagna, C. (2009). Outsourcing vs FDI in oligopoly equilibrium. *Spatial Economic Analysis, 4*(2), 149–166.
Leahy, D., & Montagna, C. (2012). Strategic investment and international outsourcing in unionised oligopoly. *Labour Economics, 19*(2), 260–269.
Leahy, D., & Pavelin, S. (2008). Playing away to win at home. *Journal of Economics and Business, 60*(5), 455–468.
Lundeen, N. (2013). Voestalpine to open direct reduction plant in Texas. *Wall Street Journal* (Europe Edition). http://online.wsj.com/article/BT-CO-20130313-702722.html. Accessed 13 Mar 2013.
Madden, P., & Pezzino, M. (2011). Oligopoly on a Salop circle with centre. *B.E. Journal of Economic Analysis & Policy, 11*(1), 1–30.
Mankiw, N. G., & Whinston, M. D. (1986). Free entry and social inefficiency. *RAND Journal of Economics, 17*(1), 48–58.
Matsumura, T., & Matsushima, N. (2003). Mixed duopoly with product differentiation: Sequential choice of location. *Australian Economic Papers, 42*(1), 18–34.
Matsumura, T., & Matsushima, N. (2004). Endogenous cost differentials between public and private enterprises: A mixed duopoly approach. *Economica, 71*(284), 671–688.
Matsushima, N., & Matsumura, T. (2003). Mixed oligopoly and spatial agglomeration. *Canadian Journal of Economics, 36*(1), 62–87.
Milliou, C. (2009). Endogenous protection of R&D investments. *Canadian Journal of Economics, 42*(1), 184–205.
Motta, M. (1993). Endogenous quality choice: Price vs. quantity competition. *Journal of Industrial Economics, 41*(2), 113–131.
Neary, J. P. (2009) *International trade in general oligopolistic equilibrium*. Working Paper, University of Oxford and CEPR.
Neary, J. P. (2010). Two and a half theories of trade. *World Economy, 33*(1), 1–19.
Neven, D. J., & Thisse, J.-F. (1990). On quality and variety competition. In *Economic decision-making: Games, econometrics and optimization: Contributions in honour of Jacques H. Dreze* (pp. 175–199). Amsterdam: North-Holland.
Oladi, R., Beladi, H., & Gilbert, J. (2007). *Strategic international outsourcing*. Working Paper.
Piga, C. A., & Poyago-Theotoky, J. (2005). Endogenous R&D spillovers and locational choice. *Regional Science and Urban Economics, 35*(2), 127–139.
Salop, S. C. (1979). Monopolistic competition with outside goods. *Bell Journal of Economics, 10* (1), 141–156.
Shaked, A., & Sutton, J. (1982) Relaxing price competition through product differentiation. *Review of Economic Studies, 49*(1), 3–13.

Spencer, B. J. (2005). International outsourcing and incomplete contracts. *Canadian Journal of Economics, 38*(4), 1107–1135.

Tobler, W. R. (1965). Computation of the correspondence of geographical patterns. *Papers of the Regional Science Association, 15*(1), 131–139.

Uchitelle, L. (2006). Goodbye, production (and maybe innovation). *New York Times*. http://www.nytimes.com/2006/12/24/business/yourmoney/24view.html?_r=0. Accessed 24 Mar 2013.

Vickers, J. (1985). Delegation and the theory of the firm. *Economic Journal, 95*(Supplement: Conference Papers), 138–147.

Voestalpine Press Release. (2013) Voestalpine constructing direct reduction plant in Texas. http://www.voestalpine.com/group/en/press/press-releases/2013-03-13-voestalpine-constructing-direct-reduction-plant-in-texas-usa.html, Accessed 13 Mar 2013.

Wu, X., & Zhang, F. (2013). *Home or overseas? An analysis of sourcing strategies under competition*. Working Paper.

Zhang, J., & Li, C. (2013). Endogenous R&D spillover and location choice in a mixed oligopoly. *Annals of Regional Science*. doi:10.1007/s00168-013-0556-2.

Zhao, K. (2012). *Delegation in a spatial game with endogenous spillovers*. Working Paper, University of Maine.

Empirical Literature on Location Choice of Multinationals

Roberto Basile and Saime Kayam

Abstract The location choices of multinational enterprises have been the center of attention both empirically and theoretically in international and regional economics during the last 20 years. Different approaches and methods have been employed to examine foreign firms' location decisions. We make a critical assessment of these approaches and their contributions to our understanding of dispersion of multinational activities across space. We start from the most influential theoretical contributions which have addressed the motivation of MNEs to be engaged in a horizontal or a vertical FDI and provide a list of the large number of foreign firms' location determinants considered in the literature. Then, we discuss the various econometric specifications used in the empirical literature to test the hypotheses on these determinants. Finally, we discuss issues for further development specifically for modeling multinationals' economic activity in space.

1 Introduction

Understanding the determinants of business location choice has traditionally been the subject of a large body of theoretical and empirical literature. Recently, there has been a growing interest in the location determinants and the spatial distribution of foreign firms operating in both manufacturing and service sectors. The focus on foreign firms' location choice (rather than on local firms) is mainly justified by three different reasons. First, the potential role of foreign firms for the economic development of a country or a region is now highly recognized. In particular, the benefits deriving from foreign firms within a country or a region

R. Basile
Department of Economics, Second University of Naples, Corso Gran Priorato di Malta, 81043 Capua (CE), Italy
e-mail: roberto.basile@unina2.it

S. Kayam (✉)
Department of Management Engineering, Istanbul Technical University, Suleyman Seba Cad. No. 90, Macka, Istanbul 34367, Turkey
e-mail: kayams@itu.edu.tr

P. Commendatore et al. (eds.), *Complexity and Geographical Economics*,
Dynamic Modeling and Econometrics in Economics and Finance 19,
DOI 10.1007/978-3-319-12805-4_13

are well known: job creation, development of subcontracting relationships with local small and medium-sized firms, introduction of new technologies, skills and capital (Castellani and Zanfei 2006). Second, as many theoretical contributions to the New Economic Geography (NEG) literature have suggested, processes of increasing regional integration, such as the ongoing enlargement of the European Union, may reshape the spatial distribution of economic activity. Since multinational enterprises (MNEs thereafter) are the main actors in the process of relocation and dislocation of economic activity, foreign firms' location behavior becomes a central topic in the empirical economic geography literature. Third, the increasing attention on foreign firms is justified by the availability of new dataset on location choices of MNEs' affiliates.

The huge amount of empirical contributions to this literature is characterized by a certain degree of heterogeneity with regards to the econometric methodology used, the choice of the dependent variable and the spatial scale adopted. Most of the advanced studies of foreign firms' location choice are now based on the *Random Utility Model* (RUM), where the industrial location decision is cast as a discrete choice problem in which profit (utility) maximizing firms select sites from a distinct set of regions. In this context, the researcher uses historical data that depict actual choices (revealed preferences) and intend to identify the factors influencing choices.[1] Within this context it is possible to distinguish two approaches. The first one focuses on the location choices made by new starting firms within a set of geographical alternatives and proceeds through discrete choice analysis. These studies use micro data (that is information on the location choice of each single foreign firm among a set of possible regional alternatives) and adopt conditional, nested or mixed logit models specified using a set of explanatory variables that intend to capture the importance of cost factors, demand variables and agglomeration economies for the business site selection process. The second approach is based on aggregated data, which count the number of new foreign entrants within each region and adopt Poisson (count) models to test the effect of the independent variables. This last approach has also found its microeconomic justification on the RUM framework, given the equivalence of the likelihood function of the conditional logit model and the Poisson regression model.

A further stream of literature uses aggregate data on inward and outward foreign direct investment (FDI) flows and stocks. These kinds of data, very popular within the international trade literature, are available also at local level for some countries (for example, for Italian provinces and for the states of USA). Linear static or dynamic panel data methods are typically used to test the effect of regional (or country) characteristics on FDI (bilateral) flows and stocks. The compatibility

[1] Some empirical research on the determinants of location choice follows the survey method. In this case, firms are required to identify the determinants of its actual location (stated preferences). The survey method allows us to obtain very rich data and to understand the ranking among alternatives, being extremely relevant when historical information is unavailable. However, the stated preferences about location may differ from the real ones, while the results are highly responsive to sample characteristics.

of this approach with the profit maximization framework (RUM) has not been clarified. Notwithstanding, within this stream of literature, we find some interesting alternative competing analytical frameworks: (1) the gravity model and (2) spatial econometric models.

In this chapter, we intend to review the empirical literature on foreign firms' location choice putting our attention on econometric methods to analyze foreign firms' location behavior. We also try to identify gaps in the literature and valuable future areas of research.

The remainder of the paper is organized as follows. In Sect. 2 we recall the most influential theoretical contributions which have addressed the motivation of MNEs to be engaged in a horizontal or a vertical FDI; we also provide a list of the large number of foreign firms' location determinants considered in the literature. Then, we discuss the various econometric specifications used in the empirical literature to test the hypotheses on these determinants (Sect. 3). Specifically, we start from discrete choice and count data models, used when foreign firms' location choices are measured by qualitative variables (dummies or counts). Then, we move to econometric specifications used to model quantitative FDI flows and stocks (Sect. 4). In this case, we start with the gravity equation and its usage in location choice literature; then we discuss recent applications of spatial econometric models to analyze FDI determinants taking spatial interaction effects (or *third country effects*) into account. Finally, we mention some issues for further development in the analysis of location choice concentrating more on the spatial approach. The chapter concludes with a short summary (Sect. 5).

2 Motivations and Types of FDI

In this section we briefly review the main economic theories developed to motivate the existence of MNEs (Sect. 2.1), then we present a classical taxonomy of FDI types (Sect. 2.2), and finally we will list the factors that attract FDIs (Sect. 2.3).

2.1 Motivations

The emergence of transnational or multinational corporations results from a number of motivations. In addressing these, Dunning (1997) proposes the OLI framework, which bases the FDI on ownership (O), location (L) and internalization (I) advantages that these firms want to pursue.[2] In the original view of Dunning, *ownership advantages* define the competitive advantages a MNE possesses *vis-à-vis*

[2]See also Dunning (1981, 1986, 1988, 2000), Dunning and Narula (1996), and Dunning et al. (1996).

the competitors in terms of tangible assets, such as scale economies or preferential access to raw materials and/or to markets. *Location advantages* stem from features of possible investment locations, such as market size, distance and labor market features, infrastructure, and identify the attractiveness of a specific location relative to others. Lastly, the *internalization advantages* arise when a firm prefers to exploit the ownership advantages in the international markets itself instead of selling, or franchising them to destination country firms.

Other approaches to the motivations of firms in engaging FDI concentrate around Krugman's (1983) proximity-concentration model (Brainard 1993) or around Markusen's (1998) knowledge-capital model. The former claims that the decision to engage in FDI is in fact optimization of the proximity to consumers and concentration of production in a location to exploit scale economies. In a comparative survey of the theories of FDI, Faeth (2009) suggests that, instead of searching for a single theory to explain FDI, a broader approach developed as a "combination of ownership advantages or agglomeration economies, market size and characteristics, cost factors, transport costs and protection and risk factors and policy variables" is more appropriate. In this regard, the knowledge-capital model provides one of the most influential explanations of the MNEs' behavior.[3]

The notion of ownership advantages is extended by Markusen (1998) to include more intangible assets such as human capital, patents and blueprints, trademarks, brand names and reputations; in a nutshell *knowledge capital*. These intangible ownership advantages have two main features, i.e. conveyability and public-good attribute within the firm. The first refers to the ease of transportation of knowledge capital to foreign affiliates with respect to physical capital, while the second has to do with the joint usage of blueprints and brand names by affiliates in various locations, once produced. Thus, the MNEs become exporters of knowledge-based assets such as managerial and engineering know-how, reputations and trademarks.

The most abstract of advantages, i.e. internalization, arises from the same public-good attribute of knowledge that creates ownership advantages. Firms prefer to establish foreign affiliates to prevent the risk of asset depletion when transferring knowledge capital through arm's length subsidiaries.

Under the assumption of plant-level scale economies, in the knowledge-capital model location advantages can be attributed to two different sources. The first one is the presence of transport costs between the market and the MNE's home country. In the absence of transport costs, the production would concentrate in a single location and all markets would be served through exports. Although transport costs are considered as the main source of location advantages in most of the models, differentiating between actual transport costs and other forms of transaction costs such as tariffs, duties or even commissions paid for bank transfers would be more beneficial in portraying the role of *distance-cum-location* accurately. A second

[3]The ideas combined in the knowledge-capital model are developed in several previous works (see, i.a., Helpman 1984, 1985; Horstmann and Markusen 1987, 1992; Ethier and Markusen 1996; Markusen and Venables 1998).

source of location advantage arises from the differences of factor intensities between source and destination countries. This is mainly valid for *cost-minimizing* MNEs or MNEs that *fragment* the production process into stages with different levels of factor intensities to exploit the factor-price differences across locations.

The general perception on location advantages recognizes that firms choose the alternative that embraces the features mostly sought by MNEs. Some of the firms may be more *cost-oriented* than others and, thus, prefer locations providing a cost advantage, be it in terms of labor, transportation or other costs. Firms with orientations such as access to technology, or access to raw materials, pick locations that would meet these needs (e.g. Buckley et al. 2007; Kumar 2007; Tolentino 2010; Zhang and Daly 2011; De Beule and Duanmu 2012; Kolstad and Wiig 2012).

More *market-oriented* firms would choose to locate in large markets (e.g. Goh and Wong 2011) or close to large markets (e.g. Chudnovsky and López 2000; Daniels et al. 2007) in order to address the need to access and/or penetrate targeted countries/regions. Maintaining export markets is another important motivation of FDI, especially from developing countries, since in most cases exports precedes outward FDI (Wells 1983). Difficulties in access to export markets such as trade barriers, access to distribution channels and consumers, and obstacles in penetrating to the markets cause firms from developing countries to become transnationals (e.g. Gang 1992; Kim and Rang 1997). Therefore, for most developing country firms, foreign investment is not a substitute for exports but a strategy to feed the export markets (e.g. Ellingsen et al. 2006).

2.2 *Types of FDI*

As discussed above, FDI flows between countries or regions are motivated by several reasons. Stylized general-equilibrium models of FDI focus, however, on market-access and cost-reduction motivations. Here, it is important to distinguish between *two-country* and *multi-country general equilibrium models*. The combination of different hypotheses (two-country vs. multi-country frameworks and market-access vs. cost-reduction motivations) gives rise to a simple taxonomy of FDI types as depicted in Table 1.

Development of formal MNE theory with a bilateral framework stems from Markusen (1984) and Helpman (1984). The first author provides a general equilibrium model where MNEs arise due to a market-access motive to substitute for export flows (when trade or tariff costs in a host country are too high), or what is termed *horizontal FDI*. The decision to undertake horizontal FDI is governed by

Table 1 Taxonomy of FDI types

Model framework	Market-access	Cost-reduction
Two-country	Horizontal FDI	Vertical FDI
Multi-country	Export-platform FDI	Vertical-specialization FDI

the *proximity-concentration trade-off* in which proximity to the host market avoids trade costs but incurs the added fixed cost of building a second production facility. If trade barriers between the parent country (where the MNE is located) and host country (where the MNE would like to make its products available) are too high, the MNE could decide to build a plant in the latter country to avoid export costs but at the expense of building a new production plant.

Alternatively, Helpman (1984) develops a general-equilibrium model where MNEs arise due to the desire to access cheaper factor inputs abroad, or what is termed *vertical FDI*. A MNE evaluates all potential destination markets to find the one that is the lowest-cost provider of the activity it wishes to relocate. MNEs will make vertical FDI if they want to access to cheaper factor inputs for their products. Both are developed in a two-country framework and have spawned significant theoretical work on MNEs.

Recent theoretical work has begun to relax the two-country assumption, leading to the development of alternative motivations for FDI. Ekholm et al. (2007), Yeaple (2003) and Bergstrand and Egger (2007) develop multi-country models of *export-platform FDI*, where a parent country invests in a particular host country with the intention of serving 'third' markets with exports of final goods from the affiliate in the host country. In this case, the motivation for FDI occurs if trade barriers between a set of destination markets are lower than trade frictions between these destination markets and the parent country. In that setup, a MNE could decide to build a plant in a host country and export to other markets. Note that using a single, well-located subsidiary provides a great deal of the proximity benefits of the pure horizontal firm without incurring additional plant-level fixed costs.

Alternatively, an MNE may set up its vertical chain of production across multiple countries to exploit the comparative advantages of various locales (Baltagi et al. 2007): *complex vertical* or *vertical specialization* FDI. Within that framework the MNE decides to split its vertical chain of production among possibly several host countries (*fragmentation*), to benefit from the comparative advantage of the hosts. Thus, complex-vertical MNE activity would be associated with exports of intermediate inputs from affiliates to third market for further (or final) processing, before being shipped to its final destination.

While both the export-platform and the complex-vertical FDI involve exports to third markets, the difference arises from the shipment of intermediate and final goods, respectively.

2.3 Factors Attracting MNEs

The factors that are expected to attract MNEs can be classified according to their relevance for the specific types of FDI, namely those that affect horizontal, vertical, export-platform or complex-vertical (see Table 2). As explained before, horizontal FDIs target the destination country market. Therefore, the location factors affecting this kind of FDIs are mainly related to market conditions. In the export-platform FDI, instead, MNEs choose a location mostly for its proximity to potential markets

Table 2 Factors that attract MNEs for different types of FDI

FDI types	Horizontal FDI	Vertical FDI	Export-platform FDI	Complex-vertical FDI
Type-specific factors	Market size, privatizations, concentration in host country, market access, distribution channels	Labor costs, infrastructure quality, human capital, natural resources, energy availability and energy costs, unionization, production costs at home country, land cost	Market potential, trade agreements, tariffs/non-tariff barriers, openness	Tariffs, non-tariff barriers, transport infrastructure quality, unionization, labor force, labor costs
Common factors	Transport costs, distance (to source country)			
Control variables	Political and economic stability, institutions, common language, legal system, free economic zones, investment promotion, corporate income taxes, government policies, international schools and hospitals, country risk, development level of banking sector, geographical characteristics (coast etc.)			

other than the host country itself. Hence, the characteristics of those potential markets, in addition to those of the destination country, are considered by MNEs. On the other hand, vertical and complex-vertical FDIs have cost minimization as their main motivation. The sources of cost minimization can arise from easier access to raw materials or cheaper labor costs with respect to the home country of the MNEs. The main difference between vertical and complex-vertical types comes from the choice of the MNE to produce the good in a single host country or in various countries, which in that case requires transport of intermediate goods to a final production location. As the number of separate locations increases, total cost of transportation incurred escalates. Therefore, in addition to other transaction costs, costs and quality of transport infrastructures become more compelling determinants under complex-vertical FDIs.

No matter the type of FDI, there are two highly impactful factors, i.e. distance and transport costs. In terms of horizontal and export-platform FDI, distance and transport costs influence the decision to invest or export while for vertical and complex-vertical FDI, these factors actually determine the production costs.

In estimating the determinants of FDI, empirical studies also control for the influence of a number of factors including political and economic stability, institutional environment, governance, etc. These control factors help the researchers to correctly identify the effect of main sources of MNE activity and henceforth the type of FDI.

3 Models for Discrete Data

In this section we discuss various econometric specifications used in the empirical literature to test the hypotheses on the determinants of foreign firms location choice. We start from discrete choice models (Sect. 3.1) and then present count data models (Sect. 3.2).

3.1 Discrete Choice Models

Recent studies on foreign firms' location choice make large use of micro datasets developed in different public and private institutions which provide information on the location decision of a large number of affiliates of MNEs. These studies usually appeal to discrete-choice models that rely on the Random Utility Maximization (RUM) framework developed by McFadden (1974).

In this context, usually multinomial, conditional, nested or mixed logit models are utilized depending on the availability of data and the research question at hand. Multinomial logit is used to model the relationship between the individual decision maker's characteristics and the likelihood of a certain choice being made. If the research question is related to the effects of the characteristics of the choice set on the decision, then conditional logit models are applied. Due to difficulty in

obtaining data on the characteristics of decision makers, numerous studies focus on the characteristics of the alternative host locations rather than the attributes of the investors and/or their perceptions of alternative locations. Therefore, such studies employ the conditional logit model of McFadden (1974) in their analysis and mainly use micro datasets that report the specific features of locations. Conditional logit and multinomial logit models have the same mathematical formulation but the former considers only the alternative-specific attributes, whereas the latter may include both individual-and alternative-specific characteristics.

In these models, a choice j made by an investor i means that the utility (or expected payoff) of the firm has been maximized with that choice. Hence, the utility obtained from j is greater than any other alternative (host country) k and these models determine the probability of j giving a higher utility then any k. Let Y_i be a random variable that indicates the choice made by i. When there are $M > 2$ alternatives or categories, each category is evaluated with respect to the reference category. If the first category is the reference then for $j = 2, \ldots, M$ the multinomial logit model estimates

$$\ln \frac{P(Y_i = j)}{P(Y_i = 1)} = \alpha_j + \sum\nolimits_{k=1}^{K} \beta_{jk} x_{ik} = Z_{ji} \tag{1}$$

and thus $(M - 1)$ equations that describe the relationship between dependent and independent variables are predicted. The predicted log odds of category j to be chosen is

$$P(Y_i = j) = \frac{exp(Z_{ji})}{1 + \sum_{h=2}^{M} exp(Z_{hi})} \tag{2}$$

and the odds for the reference category is

$$P(Y_i = 1) = \frac{1}{1 + \sum_{h=2}^{M} exp(Z_{hi})} \tag{3}$$

Studies that employ discrete choice models find factors such as agglomeration effects (Carlton 1983; Luger and Shetty 1985; Coughlin et al. 1991; Friedman et al. 1992; Wheeler and Mody 1992; Woodward 1992; Head et al. 1995; Devereux and Griffith 1998; Crozet et al. 2004), access to input and output markets (Coughlin et al. 1991; Woodward 1992; Kang and Lee 2007), factor costs especially labor costs (Crozet et al. 2004; Barrios et al. 2006; Kang and Lee 2007), transport infrastructure (Barrios et al. 2006), government policies either in terms of tax structure and tax differentials between alternative locations (Coughlin et al. 1991; Friedman et al. 1992; Woodward 1992; Devereux and Griffith 1998) or in terms of other investment promotion measures such as the free economic zones in Russia or economic zones in China (Kang and Lee 2007) to be among the most examined and decisive factors that reflect the motivations of foreign investments in terms of location choice.

Most of the early models that address the MNEs' location choice in the US are based on the characteristics of alternative states and, thus, adopt the conditional logit structure (such as Coughlin et al. 1991; Friedman et al. 1992; Head et al. 1995). Devereux and Griffith (1998) put US as the source country and examine the location choice of US firms in Europe to find agglomeration as an effective determinant as in the previously mentioned works. In fact, analogous to the US case, studies on Japanese investments reveal that agglomeration effect explains a significant part of location choice whether the location chosen is a country in Europe or China (Head and Mayer 2004; Cheng and Stough 2006). On the other hand, Blonigen et al. (2005) find that affiliation to an industrial grouping is influential on the location choice by Japanese firms. In an attempt to unfold the reasons behind the geographical dispersion of Japanese R&D activities, Shimizutani and Todo (2008) employ multinomial logit model and find that basic/applied research locates where the foreign advanced knowledge is and development/design activities locate where the market is. Investigating the potential determinants of location choice by South Korean investors in China, Kang and Lee (2007) find market size, quality of labor, transport infrastructure and government policies to be positively related to location.

Research on investment decisions in European countries have been put into spotlight by works of Crozet et al. (2004), Barrios et al. (2006) and Basile et al. (2008), which employ similar kinds of logit models in explaining location and agglomeration of multinationals in France, Ireland and eight EU countries, respectively. As data on other country firms became available, location choice of Greek (Louri et al. 2000), German and Swedish multinationals (Becker et al. 2005), French investments in Eastern and Western Europe (Disdier and Mayer 2004) were analyzed with logit models taking the alternative-specific rather than agent-specific characteristics as independent variables. In two papers examining the Portugese inward and outward FDI, Guimaraes et al. (2000) and Figueiredo et al. (2002) explore the impact of networks and social capital on inflows and of agglomeration effects on the spatial choice for foreign plants with the conditional logit model. The data used in all of these studies is location specific and do not reflect the perspectives of the managers and that is why the conditional logit not the multinomial logit model is employed.

Although outnumbered by analysis that employ conditional logit models, there are some studies, which apply multinomial logit models to explain location choice of large MNEs (Lankes and Venables 1996; Brush et al. 1999) using individual/agent-specific characteristics for various countries such as Eastern European and former Soviet Union countries (Lankes and Venables 1996), Greece (Iammarino and Pitelis 2000; Louri et al. 2000), Taiwan (Aw and Lee 2008) and Turkey (Kayam et al. 2011).

3.2 Count Data Models

Discrete choice models described so far are particularly appealing because they are obtained directly from the framework of random utility (profit) maximization and allow modeling the probability that a specific firm chooses a specific location site for its activity. In some cases, however, information on each foreign firm's location choice is not available. Rather, one might have aggregate information on the total number of new foreign firms of a specific sector entering a specific location (count data). Moreover, in practice, even when firm or establishment level data are available, the implementation of discrete choice models presents problems when one has to manage complex scenarios with a large number of spatial alternatives (for example, all municipalities or all Metropolitan Statistical Areas within a country). Finally, in some cases one needs to estimate the gap between the effective and the potential attractiveness of a region and, thus, one needs to estimate the probability that a specific location attract a certain number of foreign firms (Basile 2004; Basile et al. 2006).

Under these circumstances, the dependent variable used in the econometric analysis (the number of firms acquired or created by foreign firms operating in sector s in each region j) assumes discrete values, that is non-negative integer values (count data). The standard model for count data is the Poisson regression model, according to which the probability that the number of foreign firms operating in sector that chooses location j is n_{js} is given by:

$$P(n_{js}|\beta X_{js}) = \frac{e^{-\lambda_{js}}\lambda_{js}^{n_{js}}}{n_{js}!} \tag{4}$$

where λ_{js} is the conditional mean. Fortunately, Guimaraes et al. (2004) have demonstrated that the coefficients of the conditional logit model can be equivalently estimated by using a Poisson regression which takes n_{js} as a dependent variable and includes as explanatory variables a vector of sectoral dummy variables.[4] That is, we will obtain the same results of the conditional logit model if we admit that n_{js} follows a Poisson distribution with

$$\lambda_{js} = exp(\boldsymbol{\beta}\mathbf{X}_{js} + \boldsymbol{\theta}\,\mathbf{y}_s) \tag{5}$$

where $\mathbf{y}_s$ includes a set of sectoral dummies. However, the Poisson regression model assumes that the conditional mean λ_{js} equals the conditional variance (*equidispersion* condition). In practice, however, empirical inward FDI counts exhibit *overdispersion* and/or excess number of zeros.

[4]Recently, a large number of studies have been carried out by means of count data models (Coughlin and Segev 2000; Guimaraes et al. 2003, 2004; Figueiredo et al. 2002; Basile 2004; Basile et al. 2006; De Propris et al. 2005).

Overdispersion occurs when the variance is larger than the mean, so that the model generates consistent but inefficient estimates. In the case of location choice analysis, overdispersion is generally observed due to the concentration of foreign firms in a few areas. A way of dealing with overdispersed count data is to assume a negative binomial distribution for n_{js} which can arise as a gamma mixture of Poisson distributions.

Zero-inflation may occur when the zero outcome (that is no announcement to invest in a region) can arise from two underlying responses. On the one hand, some regions may never attract a greenfield investment, thus the outcome will be always zero. On the other hand, if the region is an attractive one, the zero outcome may be just the number of investments attracted in a given (sample) period and the response might be some positive number in a different period.

Even though negative binomial regression models capture overdispersion quite well, they are not always sufficient for modeling excess zeros. Mullahy (1986) and Lambert (1992) have addressed this problem by introducing zero-augmented models that incorporate a second model component capturing zero counts. Zero-inflation models (Lambert 1992) are mixture models that combine a count component and a point mass at zero. Hurdle models (Mullahy 1986) take a somewhat different approach and combine a left-truncated count component with a right-censored hurdle component. Examples of applications of Zero-inflated Poisson (ZIP) and Zero-inflated Negative Binomial (ZINB) models to FDI location analyses are in Tadesse and Ryan (2004), Basile (2004) and Tomlin (2000).

4 Models for FDI Flows and Stocks

More traditional studies on MNEs location decisions make use of quantitative measures of bilateral or unilateral outward FDI flows or stocks. When dyadic (i.e. bilateral) information is available, the empirical literature on the determinants of FDI deploys the services of the gravity model (Sect. 4.1), while more recent studies have used unilateral outward FDI data to apply spatial econometric techniques and test the *third-country* hypothesis (Sect. 4.2).

4.1 Gravity Models

First applied by Pöyhönen (1963) to explain international trade, the gravity model[5] rests on the conjecture that the volume of trade between two countries is directly

[5]This model is based on Newton's law of universal gravitation where the force of gravitational attraction depends directly on the masses of the objects and inversely on the distance between their centers. The formula is $F = GmM/d^2$, where F denotes the gravitational force, m and M are the masses and d is the distance between the masses. G is named as the gravitational constant.

related to the size of the economic activity indicated by their incomes and inversely related to the distance, which reflects cost of transportation, between them. Since the level of income is not sufficient to explain the purchasing power in partner economies, Linnemann (1966) suggests that population should also be taken into consideration. Thus, the traditional gravity model can be expressed as

$$T_{ij} = AY_i^{b_1} Y_j^{b_2} P_i^{b_3} P_j^{b_4} d_{ij}^{b_5} \tag{6}$$

where T_{ij} is the bilateral trade volume between countries i and j. The gravity variables (Y_i and Y_j) denote incomes, A is a constant term; P_i, P_j and d_{ij} indicate population and the distance between countries i and j, respectively. More recently, the services of the gravity model have been employed to explain bilateral trade, trade diversion or creation effects of various policies and to analyze regional integration effects (Frankel and Wei 1993; Sayan 1998; Di Mauro 2000; Feenstra et al. 2001; Nitsch 2000).

The success of the gravity model in explaining various facets of trade has attracted economists working on multinational activities and particularly on location choice of FDI (Stone and Jeon 2000; Loungani et al. 2002; Razin et al. 2003). In econometric analysis of location choice of MNEs, the log linearized form of the gravity equation estimated is given below:

$$\begin{aligned} \ln FDI_{ij,t} &= \ln A + \beta_1 \ln Y_{i,t} + \beta_2 \ln Y_{j,t} + \beta_3 \ln P_{j,t} \\ &\quad + \beta_4 \ln d_{ij} + \ln \mathbf{X}_{ij,t} \Gamma + \varepsilon_{ij,t} \end{aligned} \tag{7}$$

where $FDI_{ij,t}$ shows the FDI flows (or stocks) from source country i to destination country j at period t. The gravity variables denoted by $Y_{i,t}$, $Y_{j,t}$, $P_{j,t}$ and d_{ij} are the GDP or GDP per capita for source and destination countries, the population of the destination country at time t and the distance between source and destination countries, respectively. Other regressors capturing labor market conditions, infrastructure, institutional environment and economic stability can be included in $\mathbf{X}_{ij,t}$. Finally, $\varepsilon_{ij,t}$ is an idiosyncratic error term. Being all variables in logs, estimated $\boldsymbol{\beta}$ and Γ parameters can be interpreted as elasticities.

The gravity variables constitute the core of this approach. The parameter associated to the income of the destination country is an indicator of the type of FDI. A positive coefficient is expected for market-seeking FDI, to be showing that as income of the destination country increases then more FDI flows take place. Most common measures of income used in the literature are GDP (Brenton et al. 1999; Buch et al. 2001, 2003; Cieślik and Ryan 2004) and GDP per capita (Andreff 2002; Buch et al. 2003b).

Population is included in the gravity equation as a measure of market size, which also accounts for the purchasing power if GDP per capita is used instead of GDP as the measure of market depth. If FDIs are of the market-seeking type, then MNEs prefer to invest in a host country, where the market is large for a given level of purchasing power. Therefore, the parameter estimate is expected to be positive if

GDP per capita is used. On the other hand, if the population is large for a given level of GDP, meaning a lower purchasing power, then MNEs prefer to invest in an alternative host country. In that case, FDI is expected to decrease with population.

The distance variable, accounting for the cost of transportation, is expected to have a negative effect on trade. However, when included in the FDI gravity equation, distance has a slightly more complex impact which depends on the type of FDI (Egger 2008). If the main motivation of FDI is the market access (market seeking or horizontal FDI), then for relatively nearby destinations the MNEs could, ceteris paribus, prefer exports to FDIs; but over a certain distance-threshold, FDIs increase with the distance between source and host countries. In case of vertical FDI, MNEs actually try to decrease production costs by outsourcing parts of the production process or by acquiring cheaper raw materials, consequently distance has an increasing impact on production costs and, thence, FDI decreases with distance.

Recent studies on FDI determinants focus on transition economies and emerging markets. Most of the gravity models used in these studies employ analogous perspectives to trade studies. For example, Brenton et al. (1999) assess the relationship between trade and FDI *vis-à-vis* complementarity-substitutability and whether liberalization in Central and Eastern European Countries (CEEC) has any diversion effect on FDI to other European countries. In a similar study to trade diversion, Buch et al. (2003) address diversion of FDI from South to East European economies taking account not only of the physical distance but also of the social distance usually measured by commonality of language, history, legal system and so on. Bevan and Estrin (2004) inquire the FDI flows from West to CEECs using a gravity model enriched with transition country variables. In addition to relative market size and unit labor costs, proximity is a non-trivial factor in determining FDI flows and it influences Western European FDI in a negative fashion. Other recent country studies, which employ the gravity model to examine FDI outflows, are by Ellingsen et al. (2006) for Singapore, and by Cross et al. (2007) on China. The analysis of the outward FDI from Turkey reveal the importance of push factors for developing countries and market-seeking pattern with foreign markets being substituted for domestic market by Turkish firms (Kayam and Hisarciklilar 2009a).

4.1.1 Market Potential

As discussed above, basic gravity models are mainly oriented to test the hypothesis that FDI flows between two countries are affected by the relative size of the markets ("the mass of the objects") and by the inverse distance between the two economies. This two-country framework implies that a shock in the market size in a dyadic spatial unit ij would only affect the outcome of that dyadic unit (i.e. the bilateral FDI flow between i and j) ruling out any "third-country" effect (i.e. the effect of a shock in the market size of a country different from i and j on FDI_{ij}). As we will show in Sect. 4.2, "third-country" effects typical of a multi-country framework can be captured by using spatial econometric tools. Before that, however, it is important to introduce the notion of market potential firstly proposed by Harris (1954). He

defines market potential of an economy i ($MKTP_i$) as a weighted average of the GDP of all economies $j \neq i$:

$$MKTP_i = \sum\nolimits_{j=1} \frac{GDP_j}{d_{ij}} \tag{8}$$

where GDP_j is the GDP of the country j, which is bordering to host country i and d_{ij} is the distance between the (centroids of the) capitals of the host country i and country j, measured in kilometers. Distance should be evaluated as a proxy of transportation costs (market accessibility) as it is generally accepted in the literature (Krugman 1992).

In analyzing the market potential within the MNE context, a variety of potential measures have been used. Head and Mayer (2004) look at the determinants of agglomeration for Japanese-owned affiliates and measure potential with the demand from various locations while discounting the demand using a parameter obtained from the estimation of bilateral trade flows between the locations. Carstensen and Toubal (2004) use the region-to-region transportation costs to weigh the output of all countries in the CEE sample they consider in their quest of explaining the differences between FDI inflows to these countries. Altomonte (2002) employs three different market potential definitions: the first one is based on the interaction of the size of the neighboring markets with the degree of trade integration between countries; the second one assumes that the local markets are segmented at the country level; the last one is the traditional definition of Harris (1954). He concludes that the market accessibility is an important FDI determinant. The degree of trade integration is used to weight the market size in the standard market potential calculation instead of distance. Crozet et al. (2004) define the market potential as the sum of the local and neighboring GDPs weighted in the traditional way with inverse of the distance between locations in investigating the determinants of location choice by foreign investors in France.

Employing the potential index in a study on MNEs' location choices in the MENA region, Kayam and Hisarciklilar (2009b) expand the market potential by introducing the presence of interaction between neighboring economies. According to the authors, market potential of a country does not only stem from the size of economic activity of its neighbors, but also from interaction with those economies, i.e. trade. The economic spillover generated by a country to its neighbor will be negligible if these countries do not trade at all. Especially for the MENA region, where many countries had or still have some disputes with their neighbors, such as Israel, Qatar and Syria, Kayam and Hisarciklilar (2009b) claim that market distance does not matter when locations are segregated. They decompose the market potential index into three parts (domestic market, export and import potential indexes), making use of bilateral trade figures to measure non-domestic potential of the host country.

4.2 Spatial Econometric Models

The market potential index described above represents a means to capture spatial contagion effects of country-specific demand shocks. When included in an empirical location model, the market potential index allows us to assess the effect on the FDI attractiveness of a spatial unit i of a change (a shock) in the market size occurring not only in the same unit i, but also in all other spatial units in the system taking a distance decay effect into account. In this way, the assumption of spatial independence which characterizes many empirical analyses of FDI location choice is somehow relaxed.

More generally, in modeling FDI flows (using either unilateral or dyadic outward flows), we can include spatial interaction (or spatial dependence) effects for all the characteristics (not only the market size) of the countries or regions included in the sample. In fact, a change in the unemployment rate or in the infrastructural endowment of a region would not only influence the amount of FDI in that region, but also in all other regions in the system. However, this influence attenuates over space (distance decay effect). Failing to address this spatial dependence in the data would lead to inconsistent and inefficient parameter estimates. Spatial econometrics (Anselin 1988; LeSage and Pace 2009) provides very useful tools to model spatial dependence both in cross-section and panel data (static and dynamic) contexts.

Imagine that the objective of our analysis is to model unilateral outward FDI (for example, the outward FDI stocks from the US to all other countries in the World in a given period). In this case, a linear parametric Spatial Durbin Model (SDM) may represent the most general form to start with:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}_n\mathbf{y} + \mathbf{W}_n\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{9}$$

where $\mathbf{y}$ is the $(n \times 1)$ vector of observations of the explained variable on the n spatial units (countries or regions), $\mathbf{X}$ is a $(n \times K)$ matrix of observations of the explanatory variables on the same regions and $\boldsymbol{\varepsilon}$ is a $(n \times 1)$ vector of random shocks. Moreover, $\mathbf{W}_n$ is a $(n \times n)$ weighting matrix measuring the influence received by region i from region j. The SDM encompasses the other three spatial econometric models usually applied in the literature that is the Spatial Autoregressive Model (SAR), the Spatial Lag of X Model (SLX) and the Spatial Error Model (SEM) (see the chapter from Basile et al. 2015, in this book for more details).

A few studies have used spatial econometrics to explain the location choice of MNEs. The pioneering study in this regard is by Coughlin and Segev (2000). The authors analyze the geographic distribution of FDI within China with a spatial error model. Baltagi et al. (2007) analyze the "third-country" effects of US outward FDI in different industries to various host countries and find evidence for spatial correlation in independent variables and error terms. Garretsen and Peeters (2009) explore the third-country effects for Dutch outbound FDI. A significant and positive spatial lag coefficient implies that spatial linkages influence the location choice of

Table 3 Expected sign for spatial lag and surrounding-market potential variables

FDI motivation	Sign of spatial lag	Sign of surrounding-market potential variable
Pure horizontal	0	0
Export-platform	−	+
Pure vertical	−	0
Vertical specialization	+	0

Source: Blonigen et al. (2007)

Dutch FDI. Not only the SAR model estimated for the full sample but also the SEM adopted to the sub-samples of industry and services emphasize the impact of spatial linkages. However, the most famous and influential spatial econometric study on FDI is the one proposed by Blonigen et al. (2007). They estimate a SAR model with a market potential variable and distinguish between four different types of FDI that MNEs can undertake and can be identified based on the sign of the spatial lag parameter and of the surrounding-market potential variable (see Table 3):

- In the case of *horizontal FDI*, no spatial autocorrelation between FDI should be observed since MNEs make independent decisions about serving a market either through exports or affiliate sales. Besides, for this basic form of FDI, no market potential effect of host country should be observed since the MNE looks for access to the considered market only;
- In the case of *export-platform FDI*, as the MNE will not build a production plant in each host country, a negative spatial autocorrelation between neighboring FDI locations is expected. However, a positive effect of the surrounding-market potential variable is expected since the MNE will locate its new plant in the host country which has access to the largest surrounding market;
- In the case of *pure vertical FDI*, host countries are in competition in terms of input factor prices to receive FDI. Hence, a negative spatial autocorrelation between FDI is expected. However, since the product is shipped back to the parent country to be further processed, not any effect from surrounding-market potential is foreseen;
- In the case of the more complex form of vertical FDI (*vertical specialization*), positive spatial autocorrelation should be observed due to possible agglomeration forces such as the presence of immobile resources, since the suppliers' presence in neighboring host countries is likely to increase FDI to a particular market. However, for the same reason as in pure vertical FDI, no surrounding-market effect is predicted.

Using outbound US FDI to 35 countries over the period 1983–1998, Blonigen et al. (2007) test the dominant type of FDI which characterizes US MNEs. Even though they find a positive and significant effect of surrounding-market potential on their full sample, the authors acknowledge the fragility of their results with respect to the countries considered. Besides, they could not conclude to the presence of spatial autocorrelation for the full sample when fixed effects are included in the

specification. Garretsen and Peeters (2009) also test the dominant motivation for FDI using outward Dutch FDI to 19 countries from 1984 to 2004. When analyzing their complete sample, they not only find a positive and significant market potential effect but also positive and significant spatial autocorrelation among FDI.

The first application of spatial econometrics to the FDI in developing country context is by Kayam and Hisarciklilar (2009b), who investigate the motivation of foreign firms in the Middle East and North Africa (MENA) region by considering spatial interdependence and accounting both for horizontal/market-seeking and vertical/cost-reducing FDI. Their findings reveal that the FDI game in the MENA region is a zero-sum game both for oil and non-oil countries. Ledyaeva (2009) adopts the spatial econometric approach to examine the differences in FDI determinants in Russia between pre- and post-crises periods. She finds that spatial effects have a significant impact on FDI distribution particularly in the post-crises period. Regions in the proximity of Europe become advantageous after 1998. Kayam et al. (2013) analyze the determinants of the regional disparity in attracting FDI in Russia. The spatial distribution of FDI is investigated using regional and/or trans-regional factors. Their findings reveal that shocks in proximate regions have no effect on FDI inflows to the host region. However, FDI in a region depends on spatial lag variables such as the market size and endowment of natural resources as expected.

All in all, spatial econometric FDI models allow us to identify the "third-country" effect, that is effect of the interdependence between alternative hosts or proximate regions. The two-region framework, typical of standard gravity models, is therefore overcome in favor of a multi-region framework. However, it is worth noticing that spatial interdependence can also be introduced in a gravity model, following for example the contribution of Behrens et al. (2012) that is including the spatial lag of the dependent variable, **Wy**, on the right-hand-side of the gravity equation. This means that the FDI flow (of stock) $\mathbf{X}_{ij}$ from i to j also depends on all FDI flows (or stocks) from the other countries (regions) k to region j.

4.3 A Critical Decision in Modeling Spatial Interdependence: Weighting Matrix

A critical issue in spatial econometric analyses is the choice of the spatial weights matrix (**W**). The weights are used to position all alternative regions with respect to each other as elements of a symmetric matrix that includes all the regions in rows and in columns, i.e. the weighting matrix (**W**). In spatial econometrics analysis, the choice of **W** structure determines the way interaction between two regions is defined. There are a number of alternatives available in terms of defining **W**.

The two most extensively used symmetric weights matrices are the binary and the inverse-distance matrices. A binary symmetric spatial weights matrix assumes direct links among bordering regions (in the case of a contiguity matrix) or among regions whose distance is lower than a certain cut-off level. The neighbors are allocated

1 and all non-neighboring regions are given 0. In the inverse-distance weighting matrix, the matrix elements are the inverse distances between regions $(1/d_{ij})$. In this way, the influence of third-country characteristics declines with their distance to the host economy. A more complex spatial weights matrix is constructed by combining the full inverse-distance matrix and the binary cut-off distance matrix, i.e. defining an impact frontier (a cut-off distance) and ignoring the changes further away: the countries within this frontier would be weighted according to the distance from the target country and countries outside the frontier are not considered at all.

An inverse-distance spatial weights matrix has been adopted for example by Baltagi et al. (2007). In order to ensure robustness, they also use alternative weighting structures that coincide to various decay levels, i.e. a faster-decay with inverse of squared distances and a slower-decay using inverse of the square root of distances between alternatives. They also employ a trade-based weighting matrix, where the elements are the inverse of the averages of bilateral trade flows between alternative locations. The authors calculate these weights from the pre-estimation period data to avoid any endogeneity problem that may arise.

Trade-related weighting structures have also been used by Kayam and Hisarciklilar (2009b). Specifically, they propose two slightly different trade-based weighting matrices. The first one considers the ratio between the total volume of trade between two countries (VoT_{ij}) and the total volume of trade of country i (VoT_i) as weights:

$$w_{ij}(1) = \frac{VoT_{ij}}{VoT_i} \quad if \quad i \neq j \tag{10}$$

$$= 0 \qquad if \quad i = j \tag{11}$$

The second matrix considers distance-based neighborhood relationships in addition to the bilateral trade flows. Here, the matrix elements consist of the mean shares of neighboring countries in country i's volume of trade with all its neighbors. Hence,

$$w_{ij}(1) = \frac{VoT_{ij}}{VoTn_i} \quad if\ j\ is\ a\ neighbor\ of\ i \tag{12}$$

$$= 0 \qquad otherwise \tag{13}$$

where VoT_{ij} in the above expression denotes the volume of trade between countries i and j, and $VoTn_i$ is country i's total volume of trade with all its neighbors. If the mean share of neighbors in the host's trade volume is very small or zero as in the case of Israel, then the foreign firms will not consider these countries i and j as substitutes. Continuing with the Israel's example, foreign firms willing to supply the Israel's market will invest only in that country and nowhere else, because the neighboring countries cannot be used as an export-platform. On the other hand, if countries i and j have high bilateral trade volume then foreign firms willing to supply either or both of these markets have a choice between these two locations because either could be used as export-platform.

5 Comments and Conclusions

In this chapter we have tried to summarize some of the widely used approaches and methods that have been employed to examine foreign firms' location decisions. Our focus has mainly been on econometric methods that incorporate space either explicitly or implicitly into the investigation of MNEs' decision. In Sect. 2 we tried to give an overview of the basic motivations and types of FDI and to place them into the theoretical framework of the knowledge-capital model. Following sections reviewed the empirical methodologies used to explain the location choices of MNEs starting from discrete choice models, mainly the conditional logit model of McFadden and count data models. The gravity model, which has become the workhorse of empirical trade literature is given a considerable attention as a macro approach applied to location choice of MNEs. Either utilized together or separately with the gravity equation, the market potential models have been used to investigate the location decision for foreign investments. Therefore, we discuss the empirical studies that adopt the potential approach. There, we particularly emphasize that the standard definition of economic potential may not work in all contexts and a considerable attention has to be paid to that matter particularly when studying the location choice for horizontal FDI.

Including space into econometric analysis has been cumbersome in many areas of economics. Spatial econometrics constitute a recently acquired methodology that has the potential to dominate research on location choice of MNEs for its' convenience in exploring the impact of spatial linkages. The approaches taken in this literature are based on unilateral flows or stocks between source and destination countries. The empirical literature, as to our knowledge, has not attempted to estimate a spatial econometric model making use of dyadic data. We anticipate that as a result of such an investigation, the improvement in our understanding of the motivations and factors influencing MNEs' location decision will be notable. This should be perceived as a further issue to be developed.

Usage of dyadic data in spatial econometric models is not the only issue that is worth following. A relatively more addressed concern is the specification of the weighting or interaction matrix in spatial econometrics. Different types of weighting matrices used in location literature of FDI were given a considerable attention in the chapter. However, there are two aspects that requires further investigation. These are the choice of the weighting matrix and the conventional approach of using the same weighting matrix for both weighting the dependent and independent variables in the spatial econometric model. Above, we gave examples of context-dependent interaction matrices employed in the literature and mentioned that the studies, which choose to use separate matrices to weight dependent and independent variables. These issues, when explored in detail, has the potential of increasing the state of knowledge on not only location choice but also other space related inquiries.

Acknowledgements We are grateful to a referee that with his comments helped us to reformulate the analysis. We are responsible for any remaining errors.

Appendix: A Sample of Studies on MNEs' Location Choice

Study	Method	Countries	Findings
Altomonte (2002)	PM	CEE countries	Market potential and accessibility
Andreff (2002)	OLM	176 countries	Level of development and industry distribution in the home country
Becker et al. (2005)	CLM	German and Swedish MNEs	Market size, labor skill, labor cost, trade and investing cots
Baltagi et al. (2007)	SE	US outward FDI	Third country effects, complex FDI
Barrios et al. (2006)	NL	Ireland	Agglomeration economies, regional policy, technology level
Basile (2004)	NB, ZIP	Italian provinces	Location determinants differ according to entry mode (greenfield vs. acquisitions); Infrastructures
Basile et al. (2006)	NB	8 EU countries	Country border effects, institutions
Basile et al. (2008)	MXL	8 EU countries	Structural and cohesion funds
Bevan and Estrin (2004)	GM	12 transition economies	Unit labor costs, market size and proximity, privatization, banking sector, liberalization and institutions
Blonigen et al. (2007)	SE	US outward FDI in OECD countries	Third country effects
Brenton et al. (1999)	GM	5 EU Eastern countries	Stocks of FDI in transition countries diverge little from the expected pattern
Brush et al. (1999)	MLM	73 US, European and Japanese MNEs	National and regional characteristics
Buch et al. (2003b)	GM	Outward FDI of 7 countries	No evidence of redirection. GDP, GDP per capita, legal system, language, distance and restrictions in the recipient country
Buckley et al. (2007)	Panel	Chinese outward FDI	Market size, natural resources
Carstensen and Toubal (2004)	Dynamic panel	CEEC	Market potential, unit labor costs, labor skills, endowments, country risk, privatization

(continued)

Cheng and Stough (2006)	CLM	Japanese FDI in China	Market size, labor, land, energy cost, infrastructure, incentives, agglomeration
Cieślik and Ryan (2004)	GM	Japanese FDI in Europe	Economic potential
Coughlin and Segev (2000)	SE	China	Agglomeration economies, market size, labor supply characteristics, infrastructure
Coughlin et al. (1991)	CLM	USA	Per capita incomes, density of manufacturing activity, taxes, wages, unemployment, unionization, infrastructure, investment promotion
Crozet et al. (2004)	CLM, NL	France	Market potential
De Beule and Duanmu (2012)	CLM	Chinese and Indian acquisitions	Indian firms utilize ownership advantages. Chinese FDI is more technology-seeking
De Propris et al. (2005)	NB	Italian provinces	Agglomeration economies
Devereux and Griffith (1998)	MLM	US MNEs in Europe	Profit taxes, agglomeration economies
Disdier and Mayer (2004)	CLM, NL	French MNEs in Europe	Agglomeration economies, institutional quality
Figueiredo et al. (2002)	CLM	Outward Portuguese FDI	Agglomeration economies
Friedman et al. (1992)	CLM	Japanese and European MNEs	Access to markets, labor market conditions, state and local taxes
Garretsen and Peeters (2009)	SE	Dutch FDI in 18 countries	Third-country effects
Head and Mayer (2004)	CLM, NL	Japanese firms in Europe	Market potential
Head et al. (1995)	CLM	Japanese FDI in USA	Industry-level agglomeration
Iammarino and Pitelis (2000)	MLM	Greek FDI in Bulgaria and Romania	Labor and trade costs, proximity to EU market, investment incentives, pace of transition, technology level
Kang and Lee (2007)	CLM	South Korean FDI in China	Market size and institutions
Kim and Rang (1997)	GM	Japanese and South Korean FDI	Exports and FDI substitutability vs. complementarity; market vs. cost orientation
Kolstad and Wiig (2012)	Panel	Chinese outward FDI	Natural resources, institutions

(continued)

Lankes and Venables (1996)	MLM	Western FDI in transition countries	Progress on structural reforms
Ledyaeva (2009)	SE	FDI in Russia	Market potential
Loungani et al. (2002)	MLM	Greek firms	Borrowing capacity, labor intensity, sales growth rate, firm size, familiarity with foreign markets
Razin et al. (2003)	GM	45 countries	GDP per capita, education distance, trade, setup costs, marginal productivity
Zhang and Daly (2011)	Panel	Chinese outward FDI	Market size, natural resources, trade, growth and openness

Notes: *CLM* conditional logit model, *GM* gravity model, *MLM* multinomial logit model, *MXL* mixed logit model, *NB* negative binomial model, *NL* nested logit model, *OLM* ordered logit model, *PM* probit model, *SE* spatial econometrics, *ZIP* zero-inflated Poisson model

References

Altomonte, C. (2002). Transition bursts: Market potential and the location choices of multinational enterprises. In *29th EARIE Annual Conference of the European Association for Research in Industrial Economics*.

Andreff, W. (2002). The new multinational corporations from transition countries. *Economic Systems, 26*(4), 371–379.

Anselin, L. (1988). *Spatial econometrics: Methods and models*. Dordrecht: Kluwer Academic.

Aw, B. Y., & Lee,Y. (2008). Firm heterogeneity and location choice of Taiwanese multinationals. *Journal of International Economics, 75*(1), 167–179.

Baltagi, B. H., Egger, P., & Pfaffermayr, M. (2007). Estimating models of complex FDI: Are there third-country effects? *Journal of Econometrics, 140*(1), 260–281.

Barrios, S., Görg, H., & Strobl, E. (2006). Multinationals' location choice, agglomeration economies, and public incentives. *International Regional Science Review, 29*(1), 81–107.

Basile, R. (2004). Acquisition versus greenfield investment: The location of foreign manufacturers in Italy. *Regional Science and Urban Economics, 34*(1), 3–25.

Basile, R., Benfratello, L., & Castellani, D. (2006). Attracting foreign direct investments in Europe: Are Italian regions doomed? In *Capital accumulation, productivity and growth. Monitoring Italy* (pp. 319–354). London: Palgrave Macmillan.

Basile, R., Castellani, D., & Zanfei, A. (2008). Location choices of multinational firms in Europe: The role of EU cohesion policy. *Journal of International Economics, 74*(2), 328–340.

Basile, R., Kayam, S., Mìnguez, R., Maria, M. J., & Jesus, M. (2015). Semiparametric spatial autoregressive geoadditive models. In *Complexity and geographical economics: Topics and tools*. Berlin: Springer.

Becker, S. O., Ekholm, K., Jäckle, R., & Muendler, M.-A. (2005). Location choice and employment decisions: A comparison of German and Swedish multinationals. *Review of World Economics, 141*(4), 693–731.

Behrens, K., Ertur, C., & Koch, W. (2012). Dual gravity: Using spatial econometrics to control for multilateral resistance. *Journal of Applied Econometrics, 27*(5), 773–794.

Bergstrand, J. H., & Egger, P. (2007). A knowledge-and-physical-capital model of international trade flows, foreign direct investment, and multinational enterprises. *Journal of International Economics, 73*(2), 278–308.

Bevan, A. A., & Estrin, S. (2004). The determinants of foreign direct investment into European transition economies. *Journal of Comparative Economics, 32*(4), 775–787.

Blonigen, B. A., Davies, R. B., Waddell, G. R., & Naughton, H. T. (2007). FDI in space: Spatial autoregressive relationships in foreign direct investment. *European Economic Review, 51*(5), 1303–1325.

Blonigen, B. A., Ellis, C. J., & Fausten, D. (2005). Industrial groupings and foreign direct investment. *Journal of International Economics, 65*(1), 75–91.

Brainard, S. L. (1993). *A simple theory of multinational corporations and trade with a trade-off between proximity and concentration*. Discussion paper, National Bureau of Economic Research.

Brenton, P., Di Mauro, F., & Lücke, M. (1999). Economic integration and FDI: An empirical analysis of foreign investment in the EU and in Central and Eastern Europe. *Empirica, 26*(2), 95–121.

Brush, T. H., Marutan, C. A., & Karnani, A. (1999). The plant location decision in multinational manufacturing firms: an empirical analysis of international business and manufacturing strategy perspectives. *Production and Operations Management, 8*(2), 109–132.

Buch, C., Kleinert, J., & Toubal, F. (2003). *Determinants of German FDI: new evidence from micro-data*. Discussion Paper 09/03, Economic Research Centre of the Deutsche Bundesbank.

Buch, C., Kokta, R., & Piazolo, D. (2001). *Does the East get what would otherwise flow to the South? FDI diversion in Europe*. Kiel Working Paper.

Buch, C. M., Kokta, R. M., & Piazolo, D. (2003). Foreign direct investment in Europe: Is there redirection from the South to the East? *Journal of Comparative Economics, 31*(1), 94–109.

Buckley, P. J., Clegg, L. J., Cross, A. R., Liu, X., Voss, H., & Zheng, P. (2007). The determinants of Chinese outward foreign direct investment. *Journal of International Business Studies, 38*(4), 499–518.

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*(3), 440–449.

Carstensen, K., & Toubal, F. (2004). Foreign direct investment in Central and Eastern European countries: A dynamic panel analysis. *Journal of Comparative Economics, 32*(1), 3–22.

Castellani, D., & Zanfei, A. (2006). *Multinational firms, innovation and productivity*. Cheltenham: Edward Elgar.

Cheng, S., & Stough, R. R. (2006). Location decisions of Japanese new manufacturing plants in China: A discrete-choice analysis. *The Annals of Regional Science, 40*(2), 369–387.

Chudnovsky, D., & López, A. (2000). A third wave of FDI from developing countries: Latin American TNCs in the 1990s. *Transnational Corporations, 9*(2), 31–74.

Cieślik, A., & Ryan, M. (2004). Explaining Japanese direct investment flows into an enlarged Europe: A comparison of gravity and economic potential approaches. *Journal of the Japanese and International Economies, 18*(1), 12–37.

Coughlin, C. C., & Segev, E. (2000). Foreign direct investment in China: A spatial econometric study. *The World Economy, 23*(1), 1–23.

Coughlin, C. C., Terza, J. V., & Arromdee, V. (1991). State characteristics and the location of foreign direct investment within the United States. *The Review of Economics and Statistics, 73*(4), 675–683.

Cross, A., Buckley, P., Clegg, J., Voss, H., Rhodes, M., & Zheng, P. (2007). An econometric investigation of Chinese outward direct investment. In *Multinational enterprises and emerging challenges of the 21st century* (pp. 319–354). Cheltenham: Edward Elgar.

Crozet, M., Mayer, T., & Mucchielli, J.-L. (2004). How do firms agglomerate? A study of FDI in France. *Regional Science and Urban Economics, 34*(1), 27–54.

Daniels, J. D., Krug, J. A., & Trevino, L. (2007). Foreign direct investment from Latin America and the Caribbean. *Transnational Corporations, 16*(1), 27.

De Beule, F., & Duanmu, J.-L. (2012). Locational determinants of internationalization: A firm-level analysis of Chinese and Indian acquisitions. *European Management Journal, 30*(3), 264–277.

De Propris, L., Driffield, N., & Menghinello, S. (2005). Local industrial systems and the location of FDI in Italy. *International Journal of the Economics of Business, 12*(1), 105–121.
Devereux, M. P., & Griffith, R. (1998). Taxes and the location of production: Evidence from a panel of US multinationals. *Journal of public Economics, 68*(3), 335–367.
Di Mauro, F. (2000). *The impact of economic integration on FDI and exports: A gravity approach*. Brussels: Centre for European Policy Studies.
Disdier, A.-C., & Mayer, T. (2004). How different is Eastern Europe? Structure and determinants of location choices by French firms in Eastern and Western Europe. *Journal of Comparative Economics, 32*(2), 280–296.
Dunning, J. (1988). The eclectic paradigm of international production: A restatement and some possible extensions. *Journal of International Business Studies, 19*(1), 1–31.
Dunning, J. (1997). Trade, location of economic activity and MNE: A search for an eclectic approach. In *The International Allocation of Economic Activity*. (pp. 395–418). London: Macmillan.
Dunning, J., & Narula, R. (1996). The investment development path revisited: Some emerging issues. In *Foreign direct investment and governments* (pp. 395–418). London: Routledge.
Dunning, J. H. (1981). Explaining the international direct investment position of countries: towards a dynamic or developmental approach. *Weltwirtschaftliches Archiv, 117*(1), 30–64.
Dunning, J. H. (1986). The investment development cycle revisited. *Weltwirtschaftliches Archiv, 122*(4), 667–676.
Dunning, J. H. (2000). The eclectic paradigm as an envelope for economic and business theories of MNE activity. *International Business Review, 9*(2), 163–190.
Dunning, J. H., Van Hoesel, R., & Narula, R. (1996). Explaining the 'new' wave of outward FDI from developing countries: The case of Taiwan and Korea. *Research Memoranda, 9*, 1–25.
Egger, P. (2008). On the role of distance for outward FDI. *The Annals of Regional Science, 42*(2), 375–389.
Ekholm, K., Forslid, R., & Markusen, J. R. (2007). Export-platform foreign direct investment. *Journal of the European Economic Association, 5*(4), 776–795.
Ellingsen, G., Likumahuwa, W., & Nunnenkamp, P. (2006). Outward FDI by Singapore: a different animal? *Transnational Corporations, 15*(2), 1–40.
Ethier, W. J., & Markusen, J. R. (1996). Multinational firms, technology diffusion and trade. *Journal of International Economics, 41*(1), 1–28.
Faeth, I. (2009). Determinants of foreign direct investment-a tale of nine theoretical models. *Journal of Economic Surveys, 23*(1), 165–196.
Feenstra, R. C., Markusen, J. R., & Rose, A. K. (2001). Using the gravity equation to differentiate among alternative theories of trade. *Canadian Journal of Economics/Revue canadienne d'économique, 34*(2), 430–447.
Figueiredo, O., Guimaraes, P., & Woodward, D. (2002). Home-field advantage: Location decisions of Portuguese entrepreneurs. *Journal of Urban Economics, 52*(2), 341–361.
Frankel, J. A., & Wei, S.-J. (1993). *Trade blocs and currency blocs*, no. 4335. Cambridge, MA: National Bureau of Economic Research.
Friedman, J., Gerlowski, D. A., & Silberman, J. (1992). What attracts foreign multinational corporations? Evidence from branch plant location in the United States. *Journal of Regional Science, 32*(4), 403–418.
Gang, Y. (1992). Chinese transnational corporations. *Transnational Corporations, 1*(2), 125–133.
Garretsen, H., & Peeters, J. (2009). FDI and the relevance of spatial linkages: Do third-country effects matter for Dutch FDI? *Review of World Economics, 145*(2), 319–338.
Goh, S. K., & Wong, K. N. (2011). Malaysia's outward FDI: The effects of market size and government policy. *Journal of Policy Modeling, 33*(3), 497–510.
Guimaraes, P., O. Figueirdo, & D. Woodward (2003). A tractable approach to the firm location decision problem. *Review of Economics and Statistics, 85*(1), 201–204.
Guimaraes, P., Figueiredo, O., & Woodward, D. (2000). Agglomeration and the location of foreign direct investment in Portugal. *Journal of Urban Economics*, 47(1), 115–135.

Guimaraes, P., Figueiredo, O., & Woodward, D. (2004). Industrial location modeling: Extending the random utility framework. *Journal of Regional Science, 44*(1), 1–20.

Harris, C. D. (1954). The, market as a factor in the localization of industry in the United States. *Annals of the Association of American Geographers, 44*(4), 315–348.

Head, D., & Mayer, T. (2004). The empirics of agglomeration and trade. In *Handbook of urban and regional economics* (pp. 2609–2669). New York: North-Holland.

Head, K., Ries, J., & Swenson, D. (1995). Agglomeration benefits and location choice: Evidence from Japanese manufacturing investments in the United States. *Journal of International Economics, 38*(3), 223–247.

Helpman, E. (1984). A simple theory of international trade with multinational corporations. *The Journal of Political Economy, 92*(3), 451–471.

Helpman, E. (1985). Multinational corporations and trade structure. *The Review of Economic Studies, 52*(3), 443–457.

Horstmann, I., & Markusen, J. R. (1987). Licensing versus direct investment: A model of internalization by the multinational enterprise. *Canadian Journal of Economics, 20*(3), 464–481.

Horstmann, I. J., & Markusen, J. R. (1992). Endogenous market structures in international trade (natura facit saltum). *Journal of International Economics, 32*(1), 109–129.

Iammarino, S., & Pitelis, C. (2000). Foreign direct investment and less favoured regions: Greek FDI in Bulgaria and Romania. *Global Business Review, 1*(2), 155–171.

Kang, S. J., & Lee, H. S. (2007). The determinants of location choice of South Korean FDI in China. *Japan and the World Economy, 19*(4), 441–460.

Kayam, S., & Hisarciklilar, M. (2009a) Determinants of Turkish FDI abroad, Topics in Middle Eastern and North African Economies, electronic journal, Volume 11, Middle East Economic Association and Loyola University Chicago, http://www.luc.edu/orgs/meea/.

Kayam, S., & Hisarciklilar, M. (2009b). *Spatial determinants of FDI in the MENA region*. Paper presented at the 29th Annual Meeting of the Middle East Economic Association (ASSA) ITU-ESRC Working Paper No.09/03.

Kayam, S., Hisarcıklılar, M., & Kayalıca, Ö. (2011). *Spoilt for choice: Explaining the location choice of Turkish Transnationals*. MPRA Paper No. 39150

Kayam, S., Yabrukov, A., Hisarciklilar, M. (2013). What causes the regional disparity of FDI in Russia? A spatial analysis. *Transition Studies Review, 20*(1), 1–16.

Kim, J. D., & Rang, I. S. (1997). Outward FDI and exports: The case of South Korea and Japan. *Journal of Asian Economics, 8*(1), 39–50.

Kolstad, I., & Wiig, A. (2012). What determines Chinese outward FDI? *Journal of World Business, 47*(1), 26–34.

Krugman, P. (1992). *A dynamic spatial model*. Discussion paper, National Bureau of Economic Research.

Krugman, P. R. (1983). The new theories of international trade and the multinational enterprise. In *The Mulitnational Corporation in the 1980s*. Cambridge: MIT Press.

Kumar, N. (2007). Emerging TNCs: trends, patterns and determinants of outward FDI by Indian enterprises. *Transnational Corporations, 16*(1), 1.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34*(1), 1–14.

Lankes, H.-P., & Venables, A. J. (1996). Foreign direct investment in economic transition: the changing pattern of investments. *Economics of Transition, 4*(2), 331–347.

Ledyaeva, S. (2009). Spatial econometric analysis of foreign direct investment determinants in Russian regions. *The World Economy, 32*(4), 643–666.

LeSage, J., & Pace, K. (2009). *Introduction to spatial econometrics*. Boca Raton: CRC Press.

Linnemann, H. (1966). *An econometric study of international trade flows*. Amsterdam: North-Holland.

Loungani, P., Mody, A., & Razin, A. (2002). The Global Disconnect: The Role of Transactional Distances and Scale Economics in Gravity Equations. *Scottish Journal of Political Economy, 49*, 526–543.

Louri, H., Papanastassiou, M., & Lantouris, J. (2000). FDI in the EU periphery: A multinomial logit analysis of Greek firm strategies. *Regional Studies, 34*(5), 419–427.

Luger, M. I., & Shetty, S. (1985). Determinants of foreign plant start-ups in the United States: Lessons for policymakers in the Southeast. *Vand. J. Transnat'l L., 18*, 223.

Markusen, J. R. (1984). Multinationals, multi-plant economies, and the gains from trade. *Journal of International Economics, 16*(3), 205–226.

Markusen, J. R. (1998). Multinational firms, location and trade. *The World Economy, 21*(6), 733–756.

Markusen, J. R., & Venables, A. J. (1998). Multinational firms and the new trade theory. *Journal of international economics, 46*(2), 183–203.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour. In *Frontiers in econometrics* (pp. 395–418). New York: Academic Press.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics, 33*(3), 341–365.

Nitsch, V. (2000). National borders and international trade: Evidence from the European Union. *Canadian Journal of Economics/Revue canadienne d'économique, 33*(4), 1091–1105.

Pöyhönen, P. (1963). A tentative model for the volume of trade between countries. *Weltwirtschaftliches Archiv, Bd. 90*, 93–100.

Razin, A., Rubinstein, Y., & Sadka, E. (2003). *Which countries export FDI, and how much*? Discussion paper, National Bureau of Economic Research.

Sayan, S. (1998). The black sea economic cooperation project: A substitute for or a complement to globalization efforts in the Middle East and the Balkans? Economic Research Forum for the Arab Countries, Iran & Turkey.

Shimizutani, S., & Todo, Y. (2008). What determines overseas R&D activities? The case of Japanese multinational firms. *Research Policy, 37*(3), 530–544.

Stone, S. F., & Jeon, B. N. (2000). Foreign direct investment and trade in the Asia-Pacific region: Complementarity, distance and regional economic integration. *Journal of Economic Integration, 15*(3), 460–484.

Tadesse, B., & Ryan, M. (2004). Host market characteristics, FDI, and the FDI-trade relationship. *The Journal of International Trade & Economic Development, 13*(2), 199–229.

Tolentino, P. E. (2010). Home country macroeconomic factors and outward FDI of China and India. *Journal of International Management, 16*(2), 102–120.

Tomlin, K. M. (2000). The effects of model specification on foreign direct investment models: An application of count data models. *Southern Economic Journal, 67(2)*, 460–468.

Wells, L. T. (1983). *Third world multinationals: The rise of foreign investments from developing countries* (Vol. 1). Cambridge: MIT Press Books.

Wheeler, D., & Mody, A. (1992). International investment location decisions: The case of US firms. *Journal of International Economics*, 33(1), 57–76.

Woodward, D. P. (1992). Locational determinants of Japanese manufacturing start-ups in the United States. *Southern Economic Journal, 58(3)*, 690–708.

Yeaple, S. R. (2003). The complex integration strategies of multinationals and cross country dependencies in the structure of foreign direct investment. *Journal of International Economics, 60*(2), 293–314.

Zhang, X., & Daly, K. (2011). The determinants of China's outward foreign direct investment. *Emerging Markets Review, 12*(4), 389–398.

Spatial Interactions in Agent-Based Modeling

Marcel Ausloos, Herbert Dawid, and Ugo Merlone

Abstract Agent Based Modeling (ABM) has become a widespread approach to model complex interactions. In this chapter after briefly summarizing some features of ABM the different approaches in modeling spatial interactions are discussed.

It is stressed that agents can interact either indirectly through a shared environment and/or directly with each other. In such an approach, higher-order variables such as commodity prices, population dynamics or even institutions, are not exogenously specified but instead are seen as the results of interactions. It is highlighted in the chapter that the understanding of patterns emerging from such spatial interaction between agents is a key problem as much as their description through analytical or simulation means.

The chapter reviews different approaches for modeling agents' behavior, taking into account either explicit spatial (lattice based) structures or networks. Some emphasis is placed on recent ABM as applied to the description of the dynamics of the geographical distribution of economic activities—out of equilibrium. The Eurace@Unibi Model, an agent-based macroeconomic model with spatial structure, is used to illustrate the potential of such an approach for spatial policy analysis.

M. Ausloos
Rés. Beauvallon, rue de la Belle Jardiniére, 483/0021, 4031 Liège Angleur, Euroland, Belgium

eHumanities Group, Royal Netherlands Academy of Arts and Sciences, Joan Muyskenweg 25, 1096 CJ Amsterdam, The Netherlands
e-mail: marcel.ausloos@ulg.ac.be

H. Dawid
Department of Business Administration and Economics and Center for Mathematical Economics, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany
e-mail: hdawid@wiwi.uni-bielefeld.de

U. Merlone (✉)
Department of Psychology, Università di Torino, via Verdi 10, 10124 Torino, Italy
e-mail: ugo.merlone@unito.it

P. Commendatore et al. (eds.), *Complexity and Geographical Economics*,
Dynamic Modeling and Econometrics in Economics and Finance 19,
DOI 10.1007/978-3-319-12805-4_14

1 Agent-Based Modeling

In recent years there has been a lot of excitement about the potential of agent-based modeling (ABM). We briefly summarize the ABM approach and mention some of its applications.

In agent-based modeling, a system is modeled as a collection of autonomous decision-making entities called agents (Bonabeau 2002). Each agent individually assesses its situation and makes decisions on the basis of a set of rules.

When it comes to actual models, different approaches are proposed. For example, Axelrod proposed the KISS principle (Axelrod 1997, p. 5). This principle comes from the old army slogan, "Keep it simple, stupid" and is obviously related to the Occam's razor (Lazar 2010). When dealing with complex systems this principle is vital as, when surprising results are discovered, it is quite helpful to be confident that everything can be understood in the model that produced the surprises (Axelrod 2000). Yet, other authors have opposite views and advocate more descriptive approaches, see Edmonds and Moss (2005).

Given the level of details which can be used to model agents, this discussion is reflected also in how agents' behaviors are modeled. Several authors used data gathered in experiments, see for example Dal Forno and Merlone (2004a) and Boero et al. (2010). Recently, other approaches advocated the use of grounded theory to model agents' behavior, which enables the use of both quantitative and qualitative data, see Andrews et al. (2005), Dal Forno and Merlone (2006b, 2012), and Dawid and Harting (2012).

Nevertheless to put things in the right perspective it should be kept in mind Thorngate's "postulate of commensurate complexity", i.e. it is impossible for a theory of social behavior to be simultaneously general, accurate, and simple; as a result organizational theories inevitably have tradeoffs in their development (Thorngate 1976). To illustrate this postulate Karl Weick proposed the clock metaphor which is illustrated in Fig. 1.

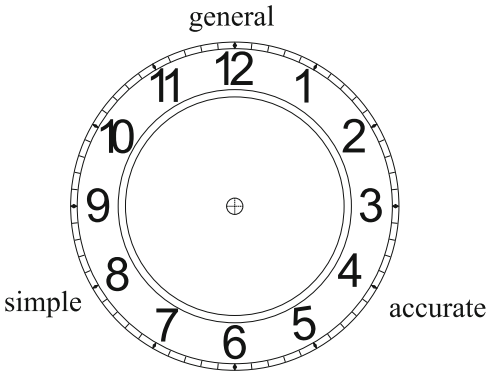


Fig. 1 Thorngate's "postulate of commensurate complexity" (Thorngate 1976) as represented by Weick's clock metaphor (Weick 1979)

This metaphor uses a clockface with *general* at 12:00, *accurate* at 4:00, and *simple* at 8:00 and shows that an explanation satisfying any two characteristics is least able to satisfy the third characteristic.

ABM has been used for theory building in social psychology (Smith and Conrey 2007), to model social processes as interactions (Macy and Willer 2002); according to Bonabeau (2002) this approach is appropriate when considering emergent phenomena in the social, political, and economic sciences. In particular in a business context, situations of interest where emergent phenomena may arise can be flow simulation, organizational simulation, market simulation, and diffusion simulation.

In agent-based modeling (ABM), a system is modeled as a collection of autonomous decision-making entities called agents which interact both with each other and with their environment. The behavior of the whole system is the result of the aggregated individual behavior of each agent. Agents can interact either indirectly through a shared environment and/or directly with each other

This way, higher-order variables such as commodity prices, population dynamics or even institutions are not specified but, instead, are the result of interaction, i.e., emergent outcomes.

2 ABM Compared to Other Approaches

As observed in O'Sullivan et al. (2012) ABMs are a relatively late arrival in fields where there is considerable previous experience with styles of model that adopt a more aggregated approach.

Also, in most fields the aggregated approach used so far continues to be quite common, therefore it would be useful to compare and assess the ABM potentialities and critical aspects, when contrasted to other approaches. In Gilbert and Troitzsch (2005) for example, ABM are compared to other social science simulation techniques in terms of communication between agents complexity of agents, number of agents and number of levels of interaction.

Related to the study of socio-economic problems implying space and time, one can find in the condensed matter literature several models of evolution, in particular modeling crystal growth. In such cases, reaction rates and different degrees of freedom coupled to their corresponding external field serve by analogy to describe the evolution of a society. Among such studies, several papers can be mentioned, as in Ausloos and Vandewalle (1996) and Vandewalle and Ausloos (1994, 1995).

In such evolutionary economics models, economic agents randomly search for new technological design by trial-and-error and run the risk of ending up in sub-optimal solutions due to interdependencies between the elements in a complex system. As argued by Frenken (1999, 2001), these models of random search are legitimate for reasons of modeling simplicity, but remain limited in scope as these models ignore the fact that agents can apply heuristics.

It has been searched within agent-based model frameworks to provide an analogy between share price instabilities and fluctuations or instabilities in electrical circuits which fluctuate in the vicinity of an unstable point (Glansdorff and Prigogine 1971), i.e., to provide the price as a thermodynamic-like variable. Recent observations have indicated that the traditional equilibrium market hypothesis (EMH; also known as Efficient Market Hypothesis) is unrealistic. It has been shown in Ausloos and Physica (2000) that the EMH is the analog of a Boltzmann equation in physics—thus having some bad properties of mean-field approximations, e.g. a Gaussian distribution of price fluctuations, rather than the empirically found "fat tails" (Mantegna and Stanley 1999). A better kinetic theory for prices can be simply derived and solved, within a Chapman-Enskog-like formalism, considering in a first approach that market agents have all identical relaxation times. In closing the set of equations, (1) an equation of state with a pressure and (2) the equilibrium (isothermal) equation for the price (taken as the order parameter) of a stock as a function of the volume of money available are obtained.

The Boltzmann kinetic equation idea has been extended in Gligor and Ignat (2002) to describe an idealized system composed by many individuals (workers, officers, business men, etc.), each of them getting a certain income and spending money for their needs. To each individual a certain time variable amount of money was associated—this defining him/her phase space coordinates. In this approximation, the exponential distribution of money in a closed economy was explicitly found. The extension of this result, including states near the equilibrium, has given the possibility to take into account the regular increase of the total amount of money, according to modern economic theories.

Other approaches consider with more detail the behavioral aspects of interaction. For example in Terna (2010) learning and adaptation are considered and interesting emerging characteristics of the agents can be observed. Also in Terna (2009) ABMs are used to understand and modify firms' behavior and to suggest possible solutions to real life applications.

3 Spatial Interactions

Spatial interaction is an important topic, see for instance Power (2009) and Stanilov (2012). In fact the interactions agents have depend on where agents are situated. Modeling the environment in which interaction takes place assume therefore an important role in modeling important phenomena. In the following we provide a rough classification of the spatial structures where interaction takes place providing examples which how flexible ABM is to deal with different spatial structures.

3.1 No Explicit Spatial Structure

In several models there is no explicit spatial structure in which the interaction among agents takes place. Rather, agents are considered being part of a unique population. This kind of approach can be considered to be a particular case of the spatial interactions that will be considered in Sects. 3.2 and 3.3. Nevertheless, even when no explicit spatial structure is considered interactions among heterogeneous agents have been analyzed in agent-based modeling. One important case is given by the the N-person prisoner's dilemma game which, according to Ostrom (2000), has come to be viewed as one of the most common representations of collective action problems among other social dilemmas. Two recent contributions (Merlone et al. 2012, 2013) analyze boundedly rational agents interactions in the N-person prisoner's dilemma using agent-based modeling to take into account of agents heterogeneity.

Other social dilemmas have been considered; for example Dal Forno and Merlone (2013) obtains simulation results of interaction in the Braess Paradox using behaviors grounded on human participants behavior.

Note that a small number of locally interacting agents seems to be the most reasonable practical case often contrary to many (theoretical) opinions. In Caram et al. (2010) the agents are supposed to interact on a given market, with some "strength" depending on their size—a peer-to-peer competition. Their evolution is governed by a set of Lotka-Volterra dynamical equations, as for prey-predator problems. The fitness of the companies thus evolve according to the value of their neighbors (in a continuum space, but through binary interactions only). The role of initial conditions knowledge is emphasized in Caram et al. (2010) following some analytical studying with a few agents (= companies). Several behaviors emerge, going from one extreme in which a few agents compete with each other, passing through oscillations, reaching some clustering state, up to the case of a "winner takes the top" state, and all others drop out. The clustering phenomenon so obtained, in market language, represents the natural segmentation into big, medium, and small "players". It can happen that the segmentation can be extreme, even of the binary type. From a socioeconomic point of view, this means that a monopolistic situation is sometimes likely. Many examples come immediately in mind, but are left for the reader to think over.

A line of research has used interactions among heterogeneous agents to model asset pricing and financial time series in markets (Cerqueti and Rotundo 2012, 2010, 2008).

3.2 Interactions on Geometrical Structures

When considering geometrical structures the line represents the simplest geometrical structure. Probably, in the bi-dimensional case the most common interaction using a geometrical structure take place on a grid. The most known

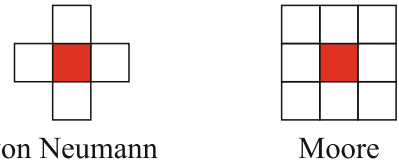


Fig. 2 Neighborhoods on bidimensional grids: von Neumann and Moore

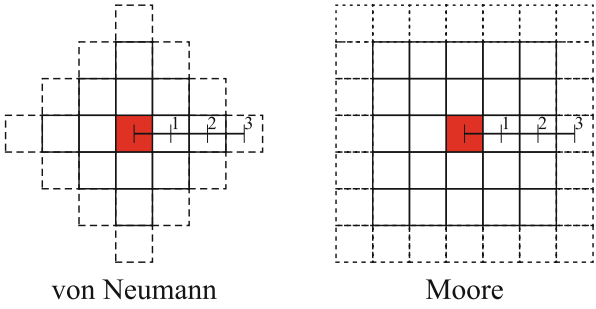


Fig. 3 von Neumann and Moore d-neighborhood with $d = 1, 2, 3$

example is given by the game of Life (Conway 1970, 1982). Actually, the game of Life is more properly a cellular automaton, that is a dynamic discrete system that can be defined as a lattice (or array) of discrete variables or "cells" that can exist in different states, Iltanen (2012). The reader may refer to Chap. 4 in North and Macal (2007) for details on how the story of agent-based modeling is related to John Conway's Game of Life and Schelling's housing segregation model (Schelling 1969).

On bi-dimensional grids several neighborhoods can be considered, the most common are the von Neumann and Moore which are illustrated in Fig. 2.

In this case interaction takes place only between adjacent cells and they can be called also von Neumann 1-neighborhood and Moore 1-neighborhood. It is possible to consider d-neighborhoods as illustrated in Fig. 3.

When the boundaries are connected a toroidal surface is obtained as illustrated in Fig. 4. In this case when either the Moore or von Neumann neighborhood has a large enough radius and the likelihood of interaction does not depend on the agents distance the geometric structure is no longer important. In this case the agents can be thought as being part of a unique population as mentioned in Sect. 3.1.

Being one of the simplest spatial interaction the one taking place on rectangular grids has been used in various applications, for a review in sociophysics see Stauffer (2003b). For example, interactions among players in two-persons game on a toroidal surface have been considered in Cerruti et al. (2005). Furthermore, in Merlone et al. (2007) finite neighborhood games have been considered with finitely many agents and with binary choices. Among two-persons games a central role is played by the Prisoner's Dilemma. For example, Nowak and May (1992) place agents on a two dimensional spatial array and observe the evolution of cooperation with deterministic players. Several interactions in organizations have the form of Prisoner's Dilemmas, therefore the modeling of such interactions has been used to explore cooperation in organizations as in Dal Forno and Merlone (2002); this

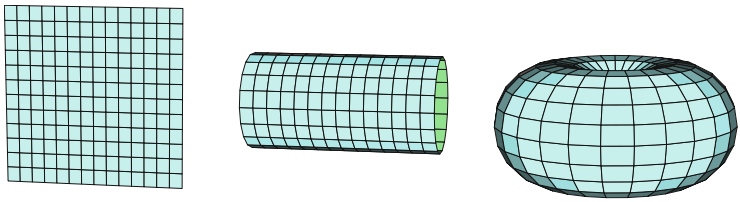


Fig. 4 Toroidal surface obtained when connecting the boundaries of a grid

analysis has been extended to consider personnel turnover in Dal Forno and Merlone (2004b, 2009a) where a reward mechanism devised to increase competition is introduced.

In Dal Forno and Merlone (2006a) the emergence of leaders is analyzed when considering interactions on a grid and distance models influence of potential leaders.

Finally, interactions on grids have been used to analyze the dynamics of industrial districts in Merlone and Terna (2006) and Merlone et al. (2008).

In Ausloos et al. (2003, 2004a,b,c), a few modern questions on economic policy: delocalization, globalization, cycles, etc., have been raised and tackled along modern statistical physics (Monte-Carlo) simulations on a rather simple but with realistic ingredient model. A highly simplified agent-based model has been first introduced in Ausloos et al. (2004a) and later developed in Ausloos et al. (2003, 2004b,c) providing a stylized geographical type of framework in order to touch upon answers on such fundamental socio-economic questions. Quantitative results similar to qualitative features are found, *a posteriori* implying that the model ingredients are close to some accepted common knowledge or stylized reality.

The model is based on agents which interact with each other under various conditions—several of them have been modified in the course of the investigations, thereby leading to several publications (Ausloos et al. 2003, 2004a,b,c).

These contributions are related to each other, but with different emphases, mainly depending on the evolution parameters. The lattice size(s), lattice symmetry (or symmetries), initial concentration(s), economic field time sequence(s), selection pressure, diffusion process rule(s), enterprise-enterprise "interaction"(s), business plan(s), number of regions, enterprise evolution law(s), and economy policy time delay implementation are all presupposed to be known for the Monte-Carlo simulation. It is found that the model even in its simplest forms can lead to a large variety of situations, including: stationary solutions and cycles, but also chaotic behavior.

The model basically consists of

1. some **space**—a square symmetry lattice—sometimes divided into three ($k = I, II, III$) **equal size** regions, connected along a horizontal axis
2. several **companies**, initially randomly placed on the lattice sites, in an
3. **environment** characterized by a real value, so called **external (exogenous), field** $F_k \in [0, 1]$,

4. and a **endogenous (internal) selection pressure** *sel*;
5. each company (i) is characterized by one real parameter $f_i \in [0, 1]$, so called its **fitness**.

The following set of actions is allowed to companies:

1. companies evolve according to their **survival probability**
   $p_i = \exp(-sel\,|f_i - F_k|)$
   compared at each time step to a predefined threshold θ, which can be time and space dependent, though most of the time has been kept constant in time and within a given region;
2. companies **may move** on the lattice one step at a time, horizontally or vertically, thus in their von Neuman neighborhood;
3. if companies meet on a site, they may
   (a) either merge with a probability b, and remain on one of the two sites which was occupied, with a new fitness value to start a "new life" with,
   (b) or create a new company with the probability $1 - b$, creating a $spin - out$ company, on some (available) site.[1]

This model is called ACP and may be described in a mean field approximation (there is no spatial structure in such an approximation) Miśkiewicz and Ausloos (2004), and Ausloos et al. (2004c) by introducing the distribution function of companies $N(t, f)$, which describes the number of companies having a given fitness f at time t. The system is then additionally characterized by the concentration of companies $c(t)$.

The ACP model (Ausloos et al. 2003, 2004b,c) looks like a reactive lattice-gas (LG) system (Simon 1993); it contains among its variants an adaptation of the Bak-Sneppen (BS) model (Bak and Sneppen 1993). The Bak-Sneppen model on a lattice has been used to model the probability of occurrence of signals commonly used in technical analysis (Rotundo and Ausloos 2007). The same model, again applied on lattices and also on scale-free networks, has been used for the analysis of the firms'dynamics (Rotundo and Scozzari 2009).

Note that the generation of new entities is more likely to occur in the case of a low concentration of companies than when this concentration is high. The merging parameter describes the reversed dependency, i.e. merging is more likely to occur in the case of a high density of companies than if the density is low.

Note also that the three region geographical-like space was introduced in order to mimic north-south or east-west problems, allowing for a third "regional continent" in the process, for the purpose of some generalization. In fact, the model can be studied in one region only. In the simulations in fact, the companies are supposed to be only in region I at first—in order to thermalize the system. Thereafter, the border between region I and region II opening looks like the Berlin wall fall, permitting the motion of firms into regions II, then III. At the same time, in some simulation, the

[1] One such site is usually available and can be chosen arbitrarily.

external field was allowed to change in region I, assuming a new value F_I, different from those in region II (F_{II}) and III (F_{III}).[2]

A few fundamental results can be outlined. One indicated that, depending on initial conditions, a cyclic process can be found, but a decay of the number of companies in some region can occur. If companies are allowed to have some strategy, based, e.g., on some information about the location, concentration or (/and) fitness of their neighbors, complex situations can occur. Interestingly it has been found that the "reaction" (or delay) time before implementing a decision (according to the strategy) induces a steady state, a cyclic state or a chaotic state. This is both interesting and frightening, since the outcome depends on the real time scale and on initial conditions (usually poorly known).

One can also observe and measure a tendency of the regional concentrations toward some equilibrium state, governed by the external field and the threshold. Interestingly, a bonus of the ACP model is that it contains parameters which can be used in scaling empirical data, like the Monte-Carlo time between two field changes—which are like changes in government policies.

An unexpected and interesting feature has been found: the local changes of the environment is leading to sharp variations, almost discontinuous ones, in the fitness and company concentrations, in particular when the field gradient is strong between regions. A lack of self-organization is thus seen at region borders.[3] One can imagine its meaning if a largest set of regions is considered.

Note that the selection pressure looks like a temperature in thermodynamics A critical value for order-disorder stares was found in Ausloos et al. (2004a). This is somewhat relevant for estimating conditions for the creation of new enterprises. Note that most of the studies pertained to conditions on the "best adapted" companies; in view of the present crisis and bankruptcy of many companies, the opposite case, could be investigated!

The ACP model has only been studied as a (Ising, spin 1/2) model. Yet, the LG approach allows already some increase in the number of degrees of freedom, characterizing the agents. However, it would be of interest to go beyond the (Ising, spin 1/2) model, and still increase the number of degrees of freedom, using real company measures (income, stocks, benefits, ...) for pursuing the investigations. Moreover the forces behind a strategy, like bargaining power and/or market threats, labor and/or transaction costs could be usefully map into endogenous fields, by raising the scalar field into a vector field, which could be space and time dependent as well. However the delay time effect and the initial conditions knowledge constraints might make the investigations rather looking like conjectures.

[2]Usually the field value was kept constant in the simulations, except for the mentioned drastic change. However some generalization is of interest, looking for relaxation and memory effects.

[3]The evolution of an economy, in which the functioning of companies is interdependent and depends on external conditions, through natural selection and somewhat random mutation is similar to bio-evolution.

Let it be observed that in contrast to the three region model presented in Commendatore and Kubin (2013), the ACP model does not consider transportation costs as an ingredient; in fact, this cost is considered to be constant in Ausloos and Physica (2000). It would be of interest to investigate further this constraint.

3.3 Interactions on Networks

More recently several ABM have been considering interaction on Networks. In Alam and Geller (2012) the relationships between agent-based social modeling and social network analysis are discussed throughly.

It is important to consider that networks in ABM can play different roles. The physical space on which the interaction takes place can be modeled by a network and obviously all the structures described in the previous sections can be represented as networks. In this case agents are not represented by the network nodes, rather they interact using the network as a physical support. For example, according to Gilbert and Troitzsch (2005), the model of segregation proposed by Schelling in Schelling (1969) can be considered as a migration model, i.e., a cellular automata where actors are not confined to a particular cell. In this case, the line where interaction takes places can be viewed also as network.

When agents are not identified with the nodes and move on physical space interacting depending on their position, a social interaction network is defined.

For example, when all nodes are connected we have a complete graph (see, Wasserman and Faust 1999), in this case the network structure becomes trivial and we are back to the cases described in Sect. 3.1.

The relationship between interactions and the network can be thought at least in two directions. On one side it is possible to assume that the pattern of interactions define the connection among nodes. For example it can be assumed that two nodes are connected if and only if at least an interaction takes place; in this case we have a dichotomous network. By contrast, when the strength of links is given by the number of interactions we have valued networks (see, Wasserman and Faust 1999).

On the other hand it can be assumed that the connections of the Network nodes are given and that interactions can take place only between connected nodes. In this case, the network defines the structure on which the interaction takes place. Finally, it is important to observe that the geometrical structures described in Sect. 3.2 can be modeled using networks. In this sense networks are a generalization of the geometrical structures so far considered.

When considering the first case, i.e., only some nodes are connected either directionally or nondirectionally it is possible to study particular interactions as those of a supervised team as in Dal Forno and Merlone (2003, 2007a, 2009c, 2012). These contributions ground agent behaviors on human participants experiments.

There are also models in which both kinds of networks are considered. For example, Dal Forno and Merlone (2007b, 2008) consider interactions between team members using two different networks. The first one is the knowledge network

which is necessary to work interaction between agents; the second network is the one which describes the work interactions which actually take place among agents. One of the interesting findings of this contribution is that starting with completely connected knowledge network, i.e., with no structure among the agents, does not allow the emergence of the more productive teams. On the other hand, a balanced expansion of the knowledge matrix is necessary for having agents working on the most productive projects. These finding have important consequences when considering social networks and the number of connections, which are related to the Dunbar's number (Dunbar 1992, 1993). In fact, according to this Author there is a cognitive limit to the number of people with whom one can maintain stable social relationships.

The network of team workers has been further examined in Dal Forno and Merlone (2014) where structural balance is introduced. In Dal Forno and Merlone (2009b) the network interaction is used to examine the role of social entrepreneurs in the emergence of cooperation. The role of social networks for the emergence of wage inequality is studied in Dawid and Gemkow (2014) using an agent-based analysis.

Interactions on different networks are considered in the literature, see Stauffer (2003a), Durán et al. (2005), Axtell (2001), Krzakala and Zdeborová (2008), Ochrombel (2001), Sousa et al. (2005), and Stauffer (2013); for a review of complex social networks the reader may refer to Vega-Redondo (2007). According to some Authors none of the standard network models fit well with sociological observations of real social networks. An interaction model grounded in social exchange theory is proposed in Pujol et al. (2005) and Hamill and Gilbert (2009) discuss how standard network models fit well with sociological observations of real social networks and proposes a new model to create a wide variety of artificial social worlds.

Even starting with an equal distribution of goods at the beginning of a closed market, on a fixed network, with free flow of goods and money, it can be shown that the market stabilizes in time (Ausloos and Pekalski 2007). This occurs faster for small markets than large ones, and even for systems in which money is steadily deduced from the market, e.g. through taxation during exchanges.[4] It has also been found that the price of goods decreases, when taxes are introduced, likely due to the less availability of money. In fact, in extreme situations, prices may not represent actual values, as is somewhat of common understanding.

Even though the model in Ausloos and Pekalski (2007) is the most simple money-goods exchange model, (there are only two "parameters": the size of the market, or in other words, the number of agents, and the initial amount of goods and money attributed to each agent), the results are somewhat indisputable, though

[4] It was found that many characteristic features are quite similar to the "no taxes" case, but again the differences are mostly seen in the distribution of wealth—the poor gets poorer and the rich gets more rich.

frightening. Any complication of the rules, thereafter, is based on "politics", but no one knows if any change in rule produces a better situation. But it can serve some of the agents instead of others. . . .

4 ABM for Modeling Geographical Distribution of Economic Activities

Understanding which factors determine the geographical distribution of economic activities and the differences in output, income, productivity or growth rates across regions is one of the most pressing issues in economics both from a theoretical and a policy perspective. Since these distributions are shaped by patterns of spatial interactions and factor flows that evolve over time, gaining a good understanding of the mechanisms responsible for the emergent outcomes asks for the consideration of dynamic spatial processes and interactions. The main body of (theory-based) economic research in this area relies on models that capture the spatial geographical structure in a very simple form and do not provide an explicit representation of the underlying dynamic processes by relying on the assumptions the economy is always in equilibrium. Most prominent in this respect is the large body of literature on new economic geography (NEG), where standard models assume a simple Core-Periphery structure and dynamics is only considered in terms of the movement of short-run equilibrium outcomes towards a long-run equilibrium (see e.g. Combes et al. 2006 and the survey on NEG in this volume). Although these approaches provide a large range of important insights into the mechanisms responsible for aggregation respectively disaggregation of economic activity they typically do not address issues like the dynamic stability of the projected long run outcome with respect to non-equilibrium adjustment processes of the economy, the effects of path dependencies in the adjustment dynamics, the effects of different kinds of spatial frictions on the goods and factor markets, the differences between short- and long run effects of policy measures on spatial distributions or implications of firm and household heterogeneities within and between regions. Agent-based spatial models allow to address such issues, but arguably the potential of an agent-based approach in this domain has not been fully exploited yet. Although several agent-based macroeconomic models with spatial structure have been developed in recent years (apart from the work discussed below, see e.g. Wolf et al. 2013) and spatial agent-based models have been used for policy analysis in agricultural economics for some time (e.g. Berger 2001; Happe et al. 2008; Filatova et al. 2009), overall there is relatively little agent-based work dealing with the dynamic processes leading to spatial distributions of activities and spatial economic policy issues. This section reviews some of the existing work, where Sect. 4.1 focuses on literature exploring general agglomeration mechanisms and Sect. 4.2 covers studies dealing with spatial policy issues. Although no claim of completeness is made for the coverage of these subsections this review makes clear that there is substantial room for more research in economic geography using an agent-based approach.

4.1 An Agent-Based Perspective on New Economic Geography Models Out of Equilibrium

In a series of papers Fowler has used agent-based simulations to address the question in how far the qualitative findings of standard New Economic Geography models can be transferred to a setting which does not assume equilibrium a priori. In Fowler (2007) he develops an agent-based model which sticks as closely as possible to the assumptions of the standard Core-Periphery model without assuming that all markets always clear and that all households and firms act optimally given current factor costs and prices (e.g. firms use mark-up pricing based on past wage costs). Furthermore, in contrast to the standard NEG approach, where the number of firms in a region is determined by the size of the local labor market, in Fowler (2007) firm exit-entry and relocation processes are explicitly modeled. The simulation results reported in Fowler (2007) show that the agent-based version of the Core-Periphery model in a large fraction of runs generates full agglomeration of workers in the long run, even for parameter constellations where the analytical version would predict relatively equal distribution of activities across regions. Furthermore, for parameter settings where also the analytical model yields full agglomeration the agent-based model in the majority of runs ends up with full agglomerations in regions different from the one predicted by the standard NEG analysis. Fowler identifies the interplay of worker and firm relocation dynamics as the source for these discrepancies. Whereas workers have incentives to move to regions where wages are highest, for firms regions are most attractive where factor costs, in particular wages, are low. This leads to rationing of firms on the labor market, regional unemployment and overall path dependencies yielding different agglomeration locations in different runs and in most cases to outcomes that are not compatible with the results of the equilibrium analysis. The conclusion in Fowler (2007) is that it remains quite unclear by which dynamic processes that are viable also out of equilibrium the equilibrium points considered in the Core-Periphery literature can be attained.

In Fowler (2011) this line of work is extended by introducing adjustment processes that allow to smooth the discrepancies between labor supply and demand in a region which arise due to the relocation dynamics described above. In particular, it is assumed that firms do not change regions but adjust employment due to hiring and firing. In addition firms might exit the market as conditions warrant and new firms might enter into a region where the gap between supply and demand is particularly high. It is shown that in cases where all workers are identical this version of the model to a large extend reproduces the results of the equilibrium analysis. However, the picture changes significantly if heterogeneity among firms and workers is introduced. In particular, in a scenario where reservation prices of workers are heterogeneous the agent-based model converges in less than 40 % to the prediction of the standard NEG Core-Periphery analysis. As pointed out in Fowler (2011) the scenarios with heterogeneous agents are however exactly the cases one might expect to see in the real world. Hence these findings suggest that substantial

additional work is needed to gain a better understanding of the dynamic process generated by the forces underlying the NEG literature. Agent-based models seem a natural tool to undertake such work.

Whereas the contributions by Fowler rely on agent-based models whose structure closely resembles that used in the NEG literature, a few other authors have studied the spatial dynamics of factor flows and economic activities based on stylized models focusing on particular types of economic activities and particular mechanisms. In Otter et al. (2001) a grid is considered on which firms and households search for locations. The focus is on the implications of the interplay of heterogeneous types of agents characterized by different decision rules governing their location search. Also the impact of changes of the radius that individuals take into account in their search is considered. An agent-base model which focuses on the interplay of learning through social interaction, creativity and location decision of workers is developed in Spencer (2012). It is analyzed how specialization between regions and the agglomeration of creative activity is influenced by different aspects like the educational system or migration incentives. In Yang and Ettema (2012) the emergence of spatial patterns of economic activity is analyzed from the perspective of firms which grow at different rates, might spawn spin-outs and relocate. The interplay of Marshall and Jacobs externalities with congestion effects is captured in an agent-based multi-level, multi-scale model. It is demonstrated how different preferences with respect to spatial proximity lead to different spatial agglomeration patterns. Finally, in the urban dynamics literature agent-based models have been developed to study agglomeration patterns on an aggregate level without capturing economic transactions on the micro-economic scale. For example, the SIMPOP model provides a rather disaggregated representation of spatial interactions by reconstructing in a multi-agent simulation model the trade flows between a large number of urban regions that are characterized by their economic portfolio (see Bretagnolle and Pumain 2010).

The use of ABM for land use has become quite popular as Matthews et al. (2007) illustrates. There are several reviews on these applications, for example see a recent review in Parker et al. (2002). In this section we will be more interested in considering the spatial interactions in terms of economic localization. Other models consider ABM for dynamic disaster environment management, see Fiedrich (2004).

4.2 Labor Flows, Regional Growth and Effects of Convergence Policy: Insights from Agent-Based Analyses

The empirical observation that regional differences in economic activity and per capita income are not only in many instances persistent over time, but may even grow due to different regional growth rates is not only of great interest for economists, but is also a major concern for policy makers. Cohesion policies aiming at the reduction of regional inequalities are among the most intensively funded

policy areas in the European Union[5] and also numerous individual countries run programs to foster the catch-up of economically lagging regions. Also, there has for a long time been a vivid political debate about the implications of regional differences with respect to institutional setups (e.g. the organization and flexibility of the labor market) and trans-regional factor flows, in particular labor and foreign direct investment, on regional growth and convergence dynamics. In spite of this large empirical importance of cohesion policy issues, systematic analyses of the short-, medium- and long-run effects of concrete policies from a theoretical perspective are largely missing. One of the reason for this lack of model-based policy analysis in this area might be that standard dynamic equilibrium models typically are too abstract to capture important characteristics of the policy measures under consideration. Also, a focus on long run steady states and balanced growth rates does not allow to study the short and medium run implications of the policies as well as potentially arising path dependencies, which seem to play an important role in empirical explanations of persistent regional differences in economic performance and policy effects. Finally, a comprehensive understanding of the effects of policy induced changes in factors like local skill endowments, labor market flexibility, labor mobility, infrastructure or support for technological development of firms must take into account the feedback between the direct effects of such measures and their implications for spatial factor flows, induced changes of firm behavior inside and outside the region as well as changes in the (distribution of) characteristics of agents (e.g. skills) in the different regions. Capturing the dynamics of these feedback requires an explicitly dynamic spatial model which includes these different sector and their connections as well as firm behavior.

4.2.1 The Eurace@Unibi Model

In a series of papers the multi-regional agent-based Eurace@Unibi model has been used to address policy questions of the kind discussed above in a dynamic spatial context. The Eurace@Unibi model is based on the agent-based macroeconomic simulation platform developed within the EU-funded EURACE project.[6] After the completion of the EURACE project in 2009 a number of authors have extended and altered the model substantially in numerous directions leading to the current version denoted as the Eurace@Unibi model. Extensive discussions of the Eurace@Unibi model can be found in Dawid et al. (2012a). In these contributions it is also shown

[5] In the period from 2007 to 2013, 347 bn Euros have been spent for cohesion policies, which makes about 36 % of the total EU budged.

[6] This project (EU IST FP6 STREP grant 035086) was carried out by a consortium lead by S. Cincotti (University of Genova), H. Dawid (University of Bielefeld), C. Deissenberg (Université de la Mediterranée), K. Erkan (TUBITAK National Research Institute of Electronics and Cryptology), M. Gallegati (Università Politecnica delle Marche), M. Holcombe (University of Sheffield), M. Marchesi (Università di Cagliari), C. Greenough (STFC—Rutherford Appleton Laboratory).

that the Eurace@Unibi model is able to reproduce a large number of empirical stylized facts on different levels of aggregation.

The model describes an economy containing labor, consumption goods ('cgoods'), capital goods (abbreviated as 'igoods' for investment goods), financial and credit markets in a regional context. The economy is inhabited by numerous instances of different types of agents: firms (consumption goods producers and capital goods producers), households and banks. Each of these agents is located in one of the regions. Additionally, there is a single central bank and a government that collects taxes and finances social benefits as well as potentially some economic policy measures, where policies might differ between regions. Finally, there is a statistical office (Eurostat) that collects data from all individual agents in the economy and generates aggregate indicators according to standard procedures. These indicators are distributed to the agents in the economy (which might use them e.g. as input for their decision rules) and also stored in order to facilitate the analysis of the simulation results. A graphical overview over the crucial parts of the model is given in Fig. 5.

Capital goods of different quality are provided by capital goods producers with infinite supply. The technological frontier (i.e. the quality of the best currently available capital good) improves over time, where technological change is driven by a stochastic (innovation) process. Firms in the consumption goods sector use capital goods combined with labor input to produce consumption goods. The labor market is populated with workers that have a finite number of general skill levels and

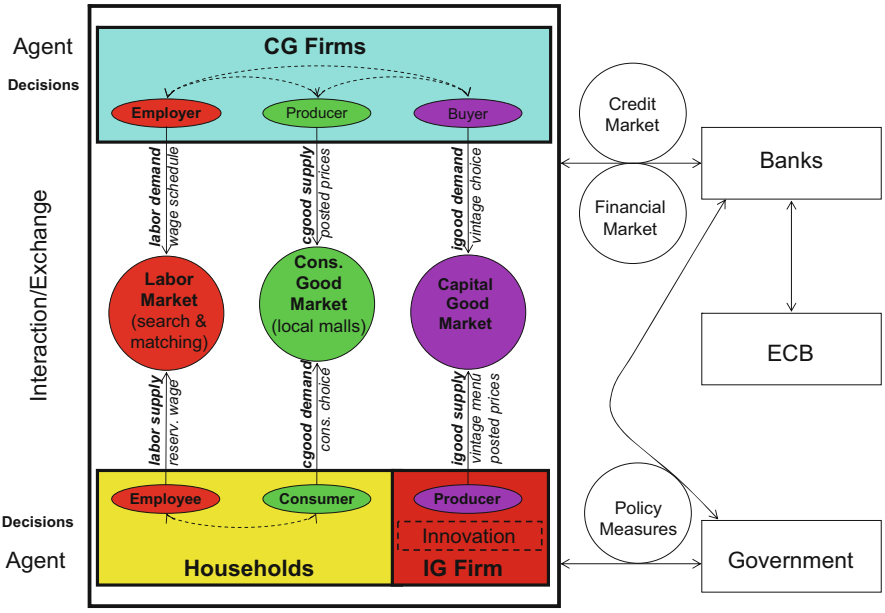


Fig. 5 Overview over the Eurace@Unibi model

acquire specific skills on the job, which they need to fully exploit the technological advantages of the capital employed in the production process. Every time when consumption goods producers invest in new capital goods they decide which quality of capital goods to select, thereby determining the speed by which new technologies spread in the economy. Consumption goods are sold at local market platforms (called malls), where firms store and offer their products and consumers come to buy goods at posted prices. Labor market interaction is described by a simple multi-round search-and-matching procedure where firms post vacancies, searching workers apply, firms make offers and workers accept/reject. Wages of workers are determined, on the one hand, by the expectation the employer has at the time of hiring about the level of specific skills of the worker, and, on the other hand, by a base wage variable, which is influenced by the (past) tightness of the labor market and determines the overall level of wages paid by a particular employer. Banks collect deposits from households and firms and give credits to firms. The interest that firms have to pay on the amount of their loan depends on the financial situation of the firm, and the amount of the loan might be restricted by the bank's liquidity and risk exposure. There is a financial market where shares of a single asset are traded, namely an index bond containing all firms in the economy. The dividend paid by each share at a certain point in time is determined by the sum of the dividends currently paid by all firms. This simple representation of a financial market is not suitable to describe speculative bubbles in the financial market, but captures important feedbacks between firm profits and households income, in a sense that fluctuations of dividends affect only the income of a particular subgroup of households, namely the owners of shares of the index bonds. The central bank provides standing facilities for the banks at a given base rate, pays interest on banks' overnight deposits and might provide fiat money to the government.

Firms that are not able to pay the financial commitments declare illiquidity. Furthermore, if at the end of the production cycle the firm has negative net worth, the firm is insolvent and insolvency bankruptcy is declared. In both cases it goes out of business, stops all productive activities and all employees loose their jobs. The firm writes off a fraction of its debt with all banks with which it has a loan and stays idle for a certain period before it becomes active again.

The spatial extensions of the markets differ. The capital goods market is global meaning that firms in all regions buy from the same global capital good producer and therefore have access to the same technologies. On the consumption goods market demand is determined locally in the sense that all consumers buy at the local mall located in their region, but supply is global because every firm might sell its products in all regional markets of the economy. Labor markets are characterized by spatial frictions determined by commuting costs that arise if workers accept jobs outside their own region. It is assumed that firms have access to all banks in the economy and, therefore, credit markets operate globally.

In contrast to dynamic equilibrium models, where it is assumed that the behavior of all actors is determined by maximization of the own (inter-temporal) objective function using correct expectations about the behavior of the other actors, agent-based simulation models need to provide explicit constructive rules that

describe how different agents build expectations and take their different decisions based on the available information, which typically does not include information about the exact structure of their economic environment. Actually, the need to provide such rules is not only based on the basic conviction underlying these models, that in most economic settings actual behavior of decision makers is far from intertemporally optimal behavior under rational expectations, but also on the fact that in most models that incorporate heterogeneity among agents and explicit interaction protocols (e.g. market rules) the characterization of dynamic equilibria is outside the scope of analytical and numerical analysis. The choice of the decision rules in the Eurace@Unibi model is based on a systematic attempt to incorporate rules that resemble empirically observable behavior documented in the relevant literature. Concerning households, this means that for example empirically identified saving rules are used and purchasing choices are described using models from the Marketing literature with strong empirical support. With respect to firm behavior the 'Management Science Approach' is followed, which aims at implementing relatively simple decision rules that match standard procedures of real world firms as described in the corresponding management literature. A more extensive discussion of the Management Science Approach can be found in Dawid and Harting (2012).

A first analysis based on the Eurace model which addresses the question how regional growth can be fostered is carried out in Dawid et al. (2008). This paper contributes to the debate whether activities to strengthen technological change should be centered on stronger regions, weaker regions, or better be uniformly distributed. The concrete policy measure under consideration is an increase in the level of general skills of workers in a region. Due to fact that higher general skills induce faster acquisition of specific skills of workers and the observation that firms can only fully exploit the quality of their physical capital stock if their workforce has appropriate specific skills (which is captured in the Eurace@Unibi model), such a policy measure should have an impact on the technology choices and the productivity of firms in a region, thereby influencing regional growth.

Under the assumption that the flow of workers between regions is hindered by substantial spatial frictions (which might be due to commuting costs or legal restrictions) the simulation experiments show that the concentration of policy measures in one region in the short run triggers stronger overall growth in the economy compared to a uniform allocation of policy measures across both regions, but that such spatial concentration of the policy effort has relatively detrimental effects on long run growth.

The findings are driven by the relatively low mobility of labor compared to that of consumption goods, which in the long run leads to an incomplete substitution of production in the low skill (less supported) region with the production in the high skill (more supported) region. We refer to Dawid et al. (2008) for a detailed discussion of the economic mechanisms underlying these results.

Subsequent work in Dawid et al. (2009) shows that the assumption of substantial spatial frictions made in Dawid et al. (2008) is indeed crucial for the qualitative findings obtained in that paper. In Dawid et al. (2009) it is demonstrated that under the empirically hardly relevant assumption of zero commuting costs (i.e. workers

are completely indifferent between working in their own or some other region as long as the same wage is offered), no significant differences between the effects of the policy types emerge. If the frictions in labor mobility are positive but small the spatially concentrated policy induces faster long-run growth than the uniform one. With a spatially concentrated policy a self-reinforcing cycle of capital and labor investments, emerges, which is triggered in the region where the policy is concentrated. The origin of this cycle is an initial asymmetry in labor costs and prices induced by the combination of a geographically concentrated skill-upgrading policy and (small) spatial labor market frictions.

The focus in Dawid et al. (2012b) is on the question how different policies of opening up labor markets accompanying an integration process of goods markets affect output and consumption dynamics in regions that start(ed) from different levels of economic development. It is explored to which extent spatial frictions with respect to labor mobility may have positive or detrimental effects on overall and region-specific variables related to the well-being of their citizens in the medium and long run.

The simulation experiments using the Eurace@Unibi model show that total output in the whole economy is lowest for closed regional labor markets. All policies that mimic an opening up of labor markets result in higher total long run output, where the differences between these policies in terms of long-run output is negligible. Effects on regional output however differ between all four policies. In particular, convergence between the regions is strongest if no labor flows between the regions are allowed. This scenario corresponds however to relatively low total output. Among the policies inducing higher total output the one generating the highest labor flows between regions reduces the inequality between regions the most.

Another concrete European regional policy issue is analyzed in Dawid et al. (2014) using the Eurace@Unibi model. The focus of the paper is on the effectiveness of different types of cohesion policies with respect to convergence of regions. Motivated by the main instruments used by the European Union (European Fund for Regional Development, European Social Fund) the effects of two types of policies are compared: technology policy, providing subsidies for firms in an economically lagging region who invest in technologies at the technological frontier, and human capital policy, inducing an improvement of the distribution of general skills in the workforce in the target region. Two different setups are considered where in the first setup the labor markets are fully integrated such that there are small frictions and all workers have almost unhindered access to both local labor markets. In the other setup the labor markets are completely separated and workers can only work in their home region.

The main results of the analysis are that the human capital policy is only effective in terms of fostering cohesion if labor markets are separated. If labor markets are integrated, output actually falls in the lagging region at which the policy is targeted. Technology policies speed up convergence for integrated and separated labor markets. The negative implications of the human capital policy under open labor markets arise although the direct goal of improving the level of specific skills

and of the vintage choice in the lagging region is reached. The negative effects of the policy for region two are due to the induced changes in the labor market tightness in that region, which have implications for wage dynamics, (relative) goods prices, demand shifts and investments.

The different analyses of spatial policy issues discussed in this subsection use the possibilities opened by the use of an agent-based simulation approach to capture how different kind of spatial frictions and spatial flows of goods and production factors affect regional economic dynamics. They capture these effects in the presence of heterogeneous firms and workers, which implies that agents are differently affected by the spatial flows and that the characteristics of the spatial flows (e.g. the skill distribution of commuting workers) depends on the distribution of agents within the regions and across regions. The discussed results are first contributions in this direction but highlight the potential of the use of agent-based models for an improved understanding of policy effects in a spatial setting.

5 Conclusion

The interest on the potentiality of ABM approach to model complex interactions is evident from the number and the quality of recent publications. This approach seem to be an important tool to model spatial inequalities evolution through time as it can take into account both the complex patterns determined by economic, geographical, institutional and social factors and the non-linearities in the decision processes of the agents. In particular parsimoniously detailed model of the interaction between the relevant actors will provide the policy makers some important tools to simulate the consequences of their decisions. Since the New Economic Geography approach, describes economic systems as very simplified spatial structures, in this chapter we provided a classification and an analysis of the spatial interaction between agents. Determining the kind of spatial interaction will be the first step to build a model to approach the uneven geographical distribution of economic activities.

References

Alam, S. J., & Geller, A. (2012). Networks in agent-based social simulation. In A. J. Heppenstall, A. T. Crooks, L. M. See, & M. Batty (Eds.) *Agent-based models of geographical systems* (pp. 199–216). Netherlands: Springer.

Andrews, C., Baptista, A. I., & Patton, S. L. W. (2005). Grounded theory and multi-agent simulation for a small firm. In T. Terano, H. Kita, T. Kaneda, K. Arai, & H. Deguchi (Eds.), *Agent-based simulation: From modeling methodologies to real-world applications* (pp. 167–181). Tokyo: Springer.

Ausloos, M. (2000). Gas-kinetic theory and Boltzmann equation of share price within an equilibrium market hypothesis and ad hoc strategy. *Physica A: Statistical Mechanics and its Applications*, *284*(1), 385.

Ausloos, M., Clippe, P., Miśkiewicz, J., & Pekalski, A. (2004). A (reactive) lattice-gas approach to economic cycles. *Physica A: Statistical Mechanics and Its Applications*, *344*(1), 1.

Ausloos, M., Clippe, P., & Pekalski, A. (2003). Simple model for the dynamics of correlations in the evolution of economic entities under varying economic conditions. *Physica A: Statistical Mechanics and Its Applications*, *324*(1), 330.

Ausloos, M., Clippe, P., & Pekalski, A. (2004a). Evolution of economic entities under heterogeneous political/environmental conditions within a Bak-Sneppen-like dynamics. *Physica A: Statistical Mechanics and its Applications*, *332*, 394.

Ausloos, M., Clippe, P., & Pekalski, A. (2004b). Model of macroeconomic evolution in stable regionally dependent economic fields. *Physica A: Statistical Mechanics and Its Applications*, *337*(1), 269.

Ausloos, M., & Pekalski, A. (2007). Model of wealth and goods dynamics in a closed market. *Physica A: Statistical Mechanics and Its Applications*, *373*, 560.

Ausloos, M., & Vandewalle, N. (1996). Growth models with internal competition. *Acta Physica Polonica Series B*, *27*, 737.

Axelrod, R. M. (1997). *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton, NJ: Princeton University Press.

Axelrod, R. M. (2000). On six advances in cooperation theory. *Analyse & Kritik*, *22*, 130.

Axtell, R. (2001). *Effects of interaction topology and activation regime in several multi-agent systems*. New York: Springer.

Bak, P., & Sneppen, K. (1993). Punctuated equilibrium and criticality in a simple model of evolution. *Physical Review Letters*, *71*(24), 4083.

Berger, T. (2001). Agent-based spatial models applied to agriculture: A simulation tool for technology diffusion, resource use changes and policy analysis. *Agricultural Economics*, *25*, 245.

Boero, R., Bravo, G., Castellani, M., & Squazzoni, F. (2010). Why bother with what others tell you? An experimental data-driven agent-based model. *Journal of Artificial Societies and Social Simulation*, *13*(3), 6.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America*, *9*(Suppl 3), 7280.

Bretagnolle, A., & Pumain, D. (2010). Simulating urban networks through multiscalar space-time dynamics (Europe and United States, 17th-20th centuries). *Urban Studies*, *47*, 2819.

Caram, L., Caiafa, C., Proto, A., & Ausloos, M. (2010). Dynamic peer-to-peer competition. *Physica A: Statistical Mechanics and Its Applications*, *389*(13), 2628.

Cerqueti, R., & Rotundo, G. (2008). Dynamics of financial time series in an inhomogeneous aggregation framework. In *Mathematical and statistical methods in insurance and finance* (pp. 67–74). New York: Springer.

Cerqueti, R., & Rotundo, G. (2010). Memory property in heterogeneously populated markets. In *Preferences and decisions* (pp. 53–67). New York: Springer.

Cerqueti, R., & Rotundo, G. (2012). The role of diversity in persistence aggregation. *International Journal of Intelligent Systems*, *27*(2), 176.

Cerruti, U., Giacobini, M., & Merlone, U. (2005). A new framework to analyze evolutionary 2×2 symmetric games. In *IEEE Proceedings of CIG'05: Symposium on Computational Intelligence and Games*, 4–6 April 2005. Colchester, Essex, UK: Essex University.

Combes, P. P., Mayer, T., & Thisse, J. F. (2006). *Economic geography: The integration of regions and nations*. Princeton: Princeton University Press.

Commendatore, P., & Kubin, I. (2013). A three-region new economic geography model in discrete time: Preliminary results on global dynamics. In G.I. Bischi, C. Chiarella, I. Sushko (Eds.) *Global analysis of dynamic models in economics and finance* (pp. 159–184). Berlin/Heidelberg: Springer.

Conway, J. H. (1970). The game of life. *Scientific American*, *223*(4), 4.

Conway, J. H. (1982). What is life. In *Winning ways for your mathematical plays* (Vol. 2, p. 927). London: Academic.

Dal Forno, A., & Merlone, U. (2002). A multi-agent simulation platform for modeling perfectly rational and bounded-rational agents in organizations. *Journal of Artificial Societies and Social Simulation*, *5*(2).

Dal Forno, A., & Merlone, U. (2003). Modular pyramidal hierarchies and social norms. An agent based model. In R. Leombruni, & M. Richiardi (Eds.) *Industry and labor dynamics* (pp. 244–255). Singapore: World Scientific

Dal Forno, A., & Merlone, U. (2004a). From classroom experiments to computer code. *Journal of Artificial Societies and Social Simulation*, *7*(3).

Dal Forno, A., & Merlone, U. (2004b). Personnel turnover in organizations: An agent-based simulation model. *Nonlinear Dynamics, Psychology, and Life Sciences*, *8*(2), 205.

Dal Forno, A., & Merlone, U. (2006a). The emergence of effective leaders: An experimental and computational approach. *Emergence: Complexity and Organization*, *8*(4), 36.

Dal Forno, A., & Merlone, U. (2006b). Building grounded agents. The lesson from Glaser and Strauss. In *Proceedings of the First World Congress on Social Simulation* (Vol. 2, pp. 377–383), Kyoto.

Dal Forno, A., & Merlone, U. (2007a). Incentives in supervised teams: An experimental and computational approach. *Journal of Social Complexity*, *3*(1), 37.

Dal Forno, A., & Merlone, U. (2007b). The evolution of coworker networks. An experimental and computational approach. In B. Edmonds, C. Hernándes, & K. G. Troitzsch (Eds.) *Social simulation technologies: Advances and new discoveries* (pp. 280–293). Hershey, NY: InformationScience Reference.

Dal Forno, A., & Merlone, U. (2008). Network dynamics when selecting work team members. In A.K. Naimzada, S. Stefani, & A. Torriero (Eds.) *Networks, topology and dynamics theory and applications to economics and social systems*. Lecture Notes in Economics and Mathematical Systems (pp. 229–240). Berlin: Springer.

Dal Forno, A., & Merlone, U. (2009a). Optimal effort in heterogeneous agents population with global and local interactions. *CUBO A Mathematical Journal*, *11*(2), 15.

Dal Forno, A., & Merlone, U. (2009b). Social entrepreneurship effects on the emergence of cooperation in networks. *Emergence: Complexity and Organization*, *11*(4), 48.

Dal Forno, A., & Merlone, U. (2009c). Individual incentives in supervised work groups: From human subject experiments to agent based simulation. *International Journal of Internet and Enterprise Management*, *6*(1), 4.

Dal Forno, A., & Merlone, U. (2012). Grounded theory based agents. In C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, & A.M. Uhrmacher (Eds.), *Proceedings of the 2012 Winter Simulation Conference*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Dal Forno A., & Merlone, U. (2014). Leaders emergence in artificial populations: the role of networks. *Quality & Quantity, 48*(5), 1853–1865

Dal Forno, A., & Merlone, U. (2013). Replicating human interaction in Braess paradox. In R. Pasupathy, S. Kim, A. Tolk, R. Hill, & M. E. Kuhl (Eds.) *Proceedings of the 2013 Winter Simulation Conference*.

Dawid, H., & Gemkow, S. (2014). How do social networks contribute to wage inequality? Insights from an agent-based analysis. *Industrial and Corporate Change, 23*, 1171–1200

Dawid, H., Gemkow, S., Harting, P., van der Hoog, S., & Neugart, M. (2012a). *Agent-based macroeconomic modeling and policy analysis: The eurace@unibi model.* Working paper, Bielefeld University.

Dawid, H., Gemkow, S., Harting, P., & Neugart, M. (2009). On the effects of skill upgrading in the presence of spatial labor market frictions: An agent-based analysis of spatial policy design. *Journal of Artifical Societies and Social Simulation*, *12*(4).

Dawid, H., Gemkow, S., Harting, P., & Neugart, M. (2012b). Labor market integration policies and the convergence of regions: The role of skills and technology diffusion. *Journal of Evolutionary Economics*, *22*, 543.

Dawid, H., Gemkow, S., Harting, P., Neugart, M., Kabus, K., & Wersching, K. (2008). Skills, innovation and growth: An agent-based policy analysis. *Jahrbücher für Nationalökonomie und Statistik/Journal of Economics and Statistics*, *228*, 251.

Dawid, H., & Harting, P. (2012). Capturing firm behavior in agent-based models of industry evolution and macroeconomic dynamics. In G. Bünsdorf (Ed.), *Applied evolutionary economics, behavior and organizations* (pp. 103–130). Cheltenham: Edward-Elgar.

Dawid, H., Harting, P., & Neugart, M. (2014). Economic Convergence: Policy Implications from a Heterogeneous Agent Model. *Journal of Economic Dynamics and Control, 44*, 54–80

Dunbar, R. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, *22*(6), 469.

Dunbar, R. I. M. (1993). Neocortical size and language. Coevolution of neocortical size, group size and language in humans. *Behavorial and Brain Sciences*, *16*, 681.

Durán, O., & Mulet, R. (2005). Evolutionary prisoners dilemma in random graphs. *Physica D: Nonlinear Phenomena*, *208*(3), 257.

Edmonds, B., & Moss, S. (2005). From kiss to kids an anti-simplistic modelling approach. In P. Davidsson, B. Logan, & K. Takadama (Eds.), *Multi-agent and multi-agent-based simulation*. Lecture Notes in Computer Science (Vol. 3415, pp. 130–144). Berlin/Heidelberg: Springer.

Fiedrich, F. (2004). *Ein high-level-architecture-basier tes multiagentensystem zur ressourcenoptimierung nach starkbeben*. Doctoral Dissertation, Universität Karlsruhe, Doctoral Dissertation.

Filatova, T., Parker, D., & van der Veen, A. (2009). Agent-based urban land markets: Agent's pricing behavior, land prices and urban land use change. *Journal of Artificial Societies and Social Simulation*, *12*(1).

Fowler, C. (2007). Taking geographical economics out of equilibrium: Implications for theory and policy. *Journal of Economic Geography*, *7*, 265.

Fowler, C. (2011). Finding equilibrium: How important is general equilibrium to the results of geographical economics? *Journal of Economic Geography*, *11*, 457.

Frenken, K. (2001). Modelling the organisation of innovative activity using the NK-model. In *Nelson and Winter Conference, Aalborg* (pp. 12–16). Citeseer.

Frenken, K., Marengo, L., & Valente, M. (1999). Interdependencies, nearly-decomposability and adaptation. In *Computational techniques for modelling learning in economics* (pp. 145–165). New York: Springer.

Gilbert, N., & Troitzsch, K. G. (2005). *Simulation for the social scientist* (2nd ed.). Milton Keynes, England: Open University Press.

Glansdorff, P., & Prigogine, I. (1971). *Structure, stabilité et fluctuations*. Paris: Masson.

Gligor, M., & Ignat, M. (2002). A kinetic approach to some quasi-linear laws of macroeconomics. *The European Physical Journal B-Condensed Matter and Complex Systems*, *30*(1), 125.

Hamill, L., & Gilbert, N. (2009). Social circles: A simple structure for agent-based social network models. *Journal of Artificial Societies and Social Simulation*, *12*(2).

Happe, K., Balmann, A., Kellermann, K., & Sahrbacher, C. (2008). Does structure matter? The impact of switching the agricultural policy regime on farm structures. *Journal of Economic Behavior and Organization*, *67*, 431.

Iltanen, S. (2012). Cellular automata in urban spatial modelling. In A. J. Heppenstall, A. T. Crooks, L. M. See, & M. Batty (Eds.) *Agent-based models of geographical systems* (pp. 69–84). Netherlands: Springer.

Krzakala, F., & Zdeborová, L. (2008). Potts glass on random graphs. *Europhysics Letters*, *81*(5), 57005.

Lazar, N. (2010). Ockham's razor. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(2), 243.

Macy, M. W., & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, *28*, 143.

Mantegna, R. N., & Stanley, H. E. (1999). *Introduction to econophysics: Correlations and complexity in finance*. Cambridge: Cambridge University Press.

Matthews, R. B., Gilbert, N. G., Roach, A., Polhill, J. G., & Gotts, N. M. (2007). Agent-based land-use models: A review of applications. *Landscape Ecology*, *22*, 1447.

Merlone, U., Sandbank, D. R., & Szidarovszky, F. (2012). Systematic approach to n-person social dilemma games: Classification and analysis. *International Game Theory Review*, *14*(3), 1.

Merlone, U., Sandbank, D. R., & Szidarovszky, F. (2013). Equilibria analysis in social dilemma games with Skinnerian agents. *Mind & Society, 12*(2), 219–233

Merlone, U., Sonnessa, M., & Terna, P. (2008). Horizontal and vertical multiple implementations in a model of industrial districts. *Journal of Artificial Societies and Social Simulations*, *11*(2)

Merlone, U., Szidarovszky, F., & Szilagyi, M. N. (2007). Finite neighborhood games with binary choices. *Mathematica Pannonica*, *18*(2), 205.

Merlone, U., & Terna, P. (2006). Population symbiotic evolution in a model of industrial districts. In R. Jean-Philippe (Ed.) *Handbook of research on nature inspired computing for economics and management* (pp. 301–316). Hershey, PA: Idea Group Inc.

Miśkiewicz, J., & Ausloos, M. (2004). A logistic map approach to economic cycles.(i). the best adapted companies. *Physica A: Statistical Mechanics and Its Applications*, *336*(1), 206.

North, M. J., & Macal, C. M. (2007). *Managing business complexity: Discovering strategic solutions with agent-based modeling and simulation*. Oxford, UK: Oxford University Press.

Nowak, M. A., & May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, *359*, 826.

Ochrombel, R. (2001). Simulation of Sznajd sociophysics model with convincing single opinions. *International Journal of Modern Physics C*, *12*(07), 1091.

Ostrom, E. (2000). Collective action and the evolution of social norms. *The Journal of Economic Perspectives*, *14*(3), 137.

O'Sullivan, D., Millington, J. J., Perry, G., & Wainwright, J. (2012). Agent-based models — Because they're worth it? In A. J. Heppenstall, A. T. Crooks, L. M. See, & M. Batty (Eds.) *Agent-based models of geographical systems* (pp. 109–123). Netherlands: Springer.

Otter, H., van der Veen, A., & de Vriend, H. (2001). ABLOoM: Location behaviour, spatial patterns and agent-based modelling. *Journal of Artificial Societies and Social Simulation*, *4*(4).

Parker, D., Manson, S., Janssen, M., Hoffmann, M., & Deadman, P. (2002). Multi-agent systems for the simulation of land-use and land-cover change: A review. *Annals Association American Geography*, *93*(2), 316.

Power, C. (2009). A spatial agent-based model of n-person prisoner's dilemma cooperation in a socio-geographic community. *Journal of Artificial Societies and Social Simulation*, *12*(1).

Pujol, J. M., Flache, A., Delgado, J., & Sangüesa, R. (2005). How can social networks ever become complex? Modelling the emergence of complex networks from local social exchanges. *Journal of Artificial Societies and Social Simulation*, *8*(4).

Rotundo, G., & Ausloos, M. (2007). Microeconomic co-evolution model for financial technical analysis signals. *Physica A: Statistical Mechanics and its Applications*, *373*, 569.

Rotundo, G., & Scozzari, A. (2009). Co-evolutive models for firms dynamics. In *Networks, topology and dynamics* (pp. 143–158). New York: Springer.

Schelling, T. C. (1969). Models of segregation. *The American Economic Review*, *59*, 488.

Simon, B. (1993). *The statistical mechanics of lattice gases* (Vol. I). Princeton, New Jersey: Princeton University Press.

Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review*, *11*, 87.

Sousa, A., Malarz, K., & Galam, S. (2005). Reshuffling spins with short range interactions: When sociophysics produces physical results. *International Journal of Modern Physics C*, *16*(10), 1507.

Spencer, G. (2012). Creative economies of scale: An agent-based model of creativity and agglomeration. *Journal of Economic Geography*, *12*, 247.

Stanilov, K. (2012). Space in agent-based models. In A. J. Heppenstall, A. T. Crooks, L. M. See, & M. Batty (Eds.) *Agent-based models of geographical systems* (pp. 253–269). Netherlands: Springer.

Stauffer, D. (2003a). Sociophysics simulations. *Computing in Science and Engineering*, *5*(3), 71.

Stauffer, D. (2003b). Sociophysics-A review of recent Monte Carlo simulations. *Fractals*, *11*, 313.

Stauffer, D. (2013). A biased review of sociophysics. *Journal of Statistical Physics*, *151*, 9.

Terna, P. (2009). The epidemic of innovation-playing around with an agent-based model. *Economics of Innovation and New Technology*, *18*(7), 707.

Terna, P. (2010). An agent-based methodological framework to simulate organizations or the quest for the enterprise: Jes and jesof, java enterprise simulator and java enterprise simulator open foundation. In E. Mollona (Ed.) *Computational analysis of firms' organisation and strategic behaviour* (pp. 247–279). New York: Routledge.

Thorngate, W. (1976). 'in general' vs. 'it depends': Some comments on the gergenschlenker debate. *Personality and Social Psychology Bulletin*, *2*(3), 404.

Vandewalle, N., & Ausloos, M. (1994). Competition between two kinds of entities in a diffusion limited aggregation process. In *Diffusion processes: Experiment, theory, simulations* (pp. 283–294). New York: Springer.

Vandewalle, N., & Ausloos, M. (1995). Evolution motivated computer models. *Annual Review of Computational Physics*, *3*, 45.

Vega-Redondo, F. (2007). *Complex social networks*. New York, NY: Cambridge University Press.

Wasserman, S., & Faust, K. (1999). *Social network analysis*. Cambridge, MA: Cambridge University Press.

Weick, K. (1979). *The social psychology of organizing*. New York, NY: McGraw-Hill.

Wolf, S., Fürst, S., Mandel, A., Lass, W., Lincke, D., Pablo-Martí, F., & Jaeger, C. (2013). A multiagent model of several economic regions. *Environmental Modelling & Software*, *44*, 25.

Yang, J. H., & Ettema, D. (2012). Modelling the emergence of spatial patterns of economic activity. *Journal of Artificial Societies and Social Simulation*, *15*(4).