University of Groningen

Models of economic geography

Knaap, Thijs

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2004

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Knaap, T. (2004). *Models of economic geography: dynamics, estimation and policy evaluation*. s.n.

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 03-06-2022

Models of Economic Geography

Dynamics, Estimation and Policy Evaluation

Thijs Knaap

Publisher: Labyrint Publications
Postbus 334
2984 AX Ridderkerk
The Netherlands
Tel: 0180-463962

Printed by: Offsetdrukkerij Ridderprint B.V., Ridderkerk.

ISBN 90-5335-025-X

Rijks*universiteit* Groningen

Models of Economic Geography:
Dynamics, Estimation and Policy Evaluation

Proefschrift

ter verkrijging van het doctoraat in de
Economische Wetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
donderdag 1 april 2004
om 16.15 uur

door

Thijs Knaap

geboren op 8 april 1972
te Oldenzaal

Promotores: Prof. dr. S. Brakman
Prof. dr. mr. C.J. Jepma

Beoordelingscommissie: Prof. dr. J. V. Henderson
Prof. dr. J.P. Neary
Prof. dr. J. Oosterhaven

Acknowledgements

Work on this thesis started in 1996, when I was appointed to the department of economics at Groningen University. The project has continued much longer than any of the original participants must have thought, but I think it has now come to a satisfying conclusion.

After using eight years to write this relatively small book, I think I owe the public some explanation about how the time on this project was spent. I left Groningen late 2000 after completing the research that was to become chapter 6 of this thesis. The manuscript still had a number of loose ends, which I intended to tie up quickly.

As these things go, however, I was soon torn between two interesting projects: while my new job took off, theoretical complications that I had not foreseen turned up in several chapters of this book. Lacking the time to fix things once and for all, I soon became the owner of an interesting collection of missed deadlines. Fortunately, time was partly on my side. I came across a new paper by Redding and Venables in the summer of 2002 in which I found the missing idea for chapter 5. Dutch speakers will appreciate the fact that this paper saved my thesis. If I add that my experience in Rotterdam allowed me to improve the general-equilibrium models in this book I hope that the long time to completion is accounted for.

As with any drawn-out project, a large number of people have been instrumental to its completion. By naming a few of them here, I run the risk of neglecting others. For this I apologize beforehand.

I thank Steven Brakman for suggesting the project and subsequently allowing it to take a number of surprising turns. Our many discussions on the subject have been interesting and taught me a thing or two about keeping an eye on the subject while cutting the right corners. I further thank Catrinus Jepma for supervising the project and tirelessly reading manuscript after manuscript, suggesting many valuable improvements along the way. Both my promotors have shown an extraordinary amount of confidence in my abilities, and I hope that I have not let them down. Simon Kuipers contributed much to the smooth start of this project and involved me in many department activities, for which I am grateful. His valuable suggestions have especially improved chapters 2 and 3.

The members of my commission have certainly behaved above and beyond the call of duty. Professor Vernon Henderson showed remarkable hospitality by letting me visit Brown University for two wonderful months while I worked on chapter 5 in early 2000. The visit allowed me access to people and data that I could not otherwise have had, while exposing me to the interesting experience of winter in Rhode Island. Professor J. Peter Neary spotted a few glaring omissions and offered a large number of suggestions that have greatly improved the manuscript in its final stages. His careful reading of the entire book is appreciated. With Jan Oosterhaven I spent the summer of 2000 on the project that led to chapter 6. Working night and day to finish our model proved a good lesson in efficiency and dedication. However, the most valuable lesson came after those warm months and showed me the value of caring for your model once the computations are done. Without Jan's neverending enthusiasm, the results in chapter 6 might not have been given the attention they deserved.

Other people have directly contributed to this project, and I thank them for their work: Gordon Hanson kindly looked at a draft of chapter 5 in its early stages. Duncan Black supplied much of the data for that chapter. Ward Romp made figure 6.2. Piet Rietveld commented extensively on the original research proposal. As usual, the blame for any harm caused by this book falls on me alone.

While a graduate student, my work was made easier by the courses and workshops offered by the Dutch network of general and quantitative economics, NAKE. I gratefully acknowledge the help I received from many of its participants.

During recent years, my colleagues at Erasmus University repeatedly allowed me to take time off in order to finish this book. At the same time, their insights into general equilibrium modelling made it much better. I therefore include Leon Bettendorf, Lans Bovenberg, Peter Broer and Ben Heijdra here and thank them for their patience.

Whether I was a student, a faculty member or someone with a thesis to hand in, the department of Economics in Groningen has always treated me well. A special mention goes to my roommates Hakan Yetkiner and Marco Haan, my colleagues in International Economics, to the best computer services department in the world, to Philipp Maier for running a marathon with me, and to Steven Brakman, Ruud Koning, Elmer Sterken and Cees Sterks for accompanying me on various other distances.

Finally, all this work would not have been possible without the support and encouragement of those I hold dear: I thank my parents for their unconditional faith in my abilities and I especially thank Marianne Scholtens, who was there from start to finish, and I promise to honor her claim to the free time that now hopefully befalls me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The purpose of this book

A remarkable thing about economic activity is how unevenly it is spread on every scale. There are vast areas in the world where almost no production takes place; then there are comparatively small areas where most of the total output is produced. But when we look inside these areas of high production, we find a similar pattern: most of the production takes place in a relatively small part of the high-producing region.

Table 1.1 will illustrate the point. The table contains two statistics that indicate the level of economic activity for different regions, each relative to the largest observation: production per area and production per area per capita. The first statistic measures the density of economic activity *per se*. Differences in this measure are partly caused by an uneven distribution of the population itself. The second measure corrects for this by dividing out the population of the region.

We look at the distribution of economic activity on three different scales, each time going down to the most productive area of the previous scale. Starting with continents, we move to countries and finally regions within a country. For brevity, we list only three or four regions at each scale, always including the most and the least productive region. Two things are notable: firstly, production per area varies enormously, as stated above. Secondly, the degree of variation does not decrease as we move to a smaller scale. There seems to be a clustering of economic production going on within the world, but also within single countries.

This tendency to cluster is the main subject of this study. It clear that an understanding of the phenomenon that is active on so many scales is important by itself, but an understanding into the patterns of agglomeration can also be useful as an instrument for advise. Economics has been called a policy science (Varian 2000), the idea being that economic theory serves mainly to gauge the possibility of economic policy. As day-to-day

Scale	Region	Production per area	Production per capita per area
Continents	High Incomes	1	1
	East Asia and Pacific	0.51	0.27
	Sub-Saharan Africa	0.06	0.09
Countries	Netherlands	1	0.95
	Belgium	0.67	1
	France	0.21	0.05
	Greece	0.08	0.12
Provinces	Zuid Holland	1	0.34
	Utrecht	0.95	1
	Groningen	0.21	0.43
	Friesland	0.09	0.16

The data on continents pertains to 1997 and is from Brakman et al. (2001), table 1.A1. Production is measured as GNP at purchasing power parity (PPP). The High Incomes countries are the United States, the European Union and Japan. Country-level data is from the World Bank (2003); it is uncorrected GDP per capita for 2002. Data on the Dutch provinces is from CBS (2002) and pertains to the year 2000.

Table 1.1: Indicators of the differences in the intensity of economic activity on different scales.

economic policy often has a distinctly regional flavor,[1] the causes behind regional economic differences deserve our attention.

In this study, we elaborate therefore on the theory and empirics of economic geography and trade. The book contains chapters on all of the different aspects of this theory, each chapter dealing with a particular question: where does the theory come from? Does it explain the clustering phenomenon that we observe? What are its predictions? We proceed by testing and applying this theoretical knowledge to real-world situations: we use econometric tests to verify if the model is an accurate description of the world, and to assess the sizes of the different effects.

The new economic geography theory that we are concerned with in this study was in the spotlight of scientific attention throughout the 1990s, after a series of theoretical breakthroughs made possible large advances in understanding. The thing that set it apart from earlier work, and earned it the dangerously perishable adjective ‘new,’ was indeed a new understanding of the smallest part of the model: the firm, and the way individual firms behave in business.

---

[1] As long as policy makers are chosen by an electorate that is defined by their place of residence, local interests will always be important.

From this change in its smallest part follow important changes in the theory's overall predictions. Some of these predictions readily fit in with known facts of life: economic production clusters together in one place instead of dispersing evenly over the land. Having a model that explained these occurrences caused considerable excitement among academic and policy-oriented economists, and generated an enormous amount of derivative research.

This seems to confirm the popular notion of economists as the rather quaint type of person who sees something work in practice and then wants to know if it would work in principle.[2] Why indeed should we worry about the theoretical explanations of things like cities and industrialized countries if we already have them around? Things are difficult enough to analyze without having to explain how it all came together in the first place. And so indeed a lot of useful theory has been made *given* the existence technical progress, *given* the existence of a large city, or *given* the existence of a rich and a poor trading partner.

But such theories are necessarily incomplete. If we do not know why the prosperous region formed in the first place, we have no idea what will happen to it during the rest of its lifecycle, or what will happen if we change something in the surrounding environment. Taking certain facts as given, we are holding constant things that might change, in unpredictable ways and at uncertain times. Thus the new theories of economic geography and economic growth gave to economists an understanding of the dynamics of regions, gave them insight in the stability of the current equilibrium and showed them how it could be influenced. Furthermore, table 1.1 illustrates that 'regions' may mean quite large areas. It was these things that drew the attention of the profession to the new theories.

In this book, we will see the development of a new piece of economic theory along all of its methodological stages: we explain the principles that govern the model's properties, and determine its possible outcomes. We look for similar empirical stylized facts and use them as a first test for the model. After that, we turn to the more rigorous method of statistical analysis, to test the model and find its key parameters. Finally, we use the theory in a policy evaluation exercise, in which we try to assess the effects of a change in the economic environment.

1.2 The final frontier

The history of spatial economics presents an interesting case of selective blindness. For many years, very little attention was paid to the field by a vast majority of the economic profession. Krugman (1998) reports that

---

[2] This golden classic of any collection of Economist Jokes appears to originate from a footnote in Goldfeld (1984).

there exist *no* references to the spatial economics in any of the (then current) major economics textbooks, for instance.

Presumably, the reason for this strange absence is that according to mainstream economic theory, space is hardly relevant. Much of this conviction is caused by an assumption that is usually made in neoclassical economics, the assumption of constant returns to scale. Constant returns imply that any economic process can be split into parts, each of which is a perfect, scaled down copy of the original. That means that the efficiency of production does not change with its scale. As such, it is easy to see that location is irrelevant: even if the production process is spread across the country, according to the theory it can be sliced into as many small versions as desired and be dispersed over the land without losing efficiency. Backyard production of the whole consumption bundle is possible, and where firms and people live is completely undetermined.

This result is an element of the spatial impossibility theorem (Starrett 1978). The theorem states that a model with mobile agents on a closed, homogeneous space, who employ a constant-returns production technology, can never explain the occurrence of agglomerations.

An example of this state of affairs is the way in which international trade is modelled in the neoclassical Heckscher-Ohlin framework. The theory states that in a situation where two countries have different endowments of two production factors, each will specialize in the technology that uses the locally abundant factor intensively. That is, if country N has a lot of capital while country S is abundant in labor, international trade will take the form of N trading cars for S's agricultural products, for instance. As Krugman (1995b) shows, in this model international trade allows the world economy to produce as if there were no borders: by bringing production to where the production factors are, the 'punishment' of borders is conveniently escaped.

In this theory of international trade, spatial issues are completely absent. Countries are seen as points without a spatial dimension, because the location of production is irrelevant; as long as trade is free, we can be sure that the optimal organization of production is established.

Meanwhile, the casual observer will notice that the real world consists of many places almost devoid of human activity, and a few spots where very many people have chosen to live and work. As we noticed above, the propensity of people to cluster can be seen on many levels: in villages within a region, in the downtown area of a metropolis, or in countries within a larger union. It is apparent even from space, as figure 1.1 shows.

For many spatial economists to whom the occurrence of clustering was a natural fact, it also was natural to start their theories from the assumption of a city. The rich German tradition in spatial modelling (a well-known example is von Thünen 1842) shows that many useful things can be understood, given that a city exists. These theories were not part of mainstream

Figure 1.1: A nighttime picture of London, taken from the International Space Station. Image courtesy of the Earth Sciences and Image Analysis Laboratory, NASA Johnson Space Center (2003).

economics as it was taught in the textbooks, but they were found and used by those who wished to analyze spatial matters. The development of spatial theories continued within its own subfield for decades. Blaug (1996) writes

> Spatial economics, and particularly the theory of the location of economic activity, flourished and matured throughout the nineteenth century but in almost total isolation of mainstream economics, whether classical or neo-classical. Indeed, it is not too much to say that the whole of mainstream economics was until 1950 effectively confined to the analysis of an economic world without spatial dimensions. (p. 596)

This state of affairs continued until a theoretical breakthrough occurred that allowed economists to relinquish the assumption of constant returns. The Monopolistic Competition revolution, as it has been called (Brakman and Heijdra 2003), made it possible construct a model in which firms of a fixed, efficient, size exist.

The consequences of this new model were uncovered in a number of phases. First up was trade theory (see for instance chapter 9 of Dixit and Norman 1980, or the model in Krugman 1979), where it was found that monopolistically competitive firms, in the same sector but with differentiated products, play an important role in intra-industry trade. That is, international trade was no longer just concerned with restoring the efficient method of world production by allowing countries to specialize, but also involved the trade in product varieties of similar factor intensities. Demand for a variety of goods comes from consumers as well as firms (Ethier 1982), causing international trade in intermediate goods.

Next up, growth theory (see Romer 1986, Rebelo 1991) was expanded with endogenous growth theories, in which the decreasing returns to accumulable resources make way for constant returns. It was found that the factor behind growth was not only a growing capital stock, but also new ideas for new firms. This enabled theorists to predict the rate of growth of an economy, and make observations on the factors behind it.

Finally (for now), it was realized that with the introduction of firms that can no longer be split into pieces, spatial considerations become important. If a firm of fixed size is the efficient unit of production, where that firm will locate is a decision of some interest. Furthermore, if these firms somehow complement each other, agglomeration could be explained by a desire to be close to other firms. Thus started new economic geography (Krugman 1991b), which brought outside attention to spatial economics that was both welcomed and resented.[3]

With the introduction of mainstream methods into spatial economics comes a number of tools that are very useful in policy analysis. Explicit welfare analysis is one of them, and it allows policy makers to assess the total effect of changes in the economic environment on consumers and producers. The explicit behavioral assumptions embodied in the theory's microeconomic foundations also allow for a consistent estimation of the model's parameters, which makes research results quantifiable.

This book uses the new methods of economic geography and economic growth to add to the theory on spatial clustering and regional evolutions. We also use the new instruments to assess a policy proposal in which a new railway link is constructed between two Dutch regions. Before we get into theory and policy evaluation, we spend some time to understand the background of the theoretical breakthrough that makes the new theory possible.

1.3 Outline

We survey the literature on monopolistic competition, economic growth and location theories in chapter 2. The factor that unites these seemingly disparate fields is the concept of complementarity. We show that complementarity between firms allows for economic growth, and attracts producers to clusters of firms. There exist several ways in which complementarity between firms can assert itself, each mechanism leading to a different model of economic geography. The forces of growth and spatial clustering show their combined (and interfering) effects in theories of regional growth.

---

[3] A history of the field as it is seen by the newly arrived theorists can be read in chapters 2 and 3 of Fujita et al. (1999). The added value of their new theory receives a critical inspection in Neary (2001).

In chapter 3, we elaborate on the model in which agglomeration is caused by a link between firms, which use each other's product as an intermediate input. We make the obvious extension of firms in different sectors with an input-output matrix between them. However, we assume a sector-structure that, in analogy to the 'continuous firms' concept, can be thought of as fluid: there are no discrete groups of firms, with each group forming its own sector. Rather, we allow for maximal flexibility and do not constrain the aggregate of products that each firm uses as its intermediate input. This model can be used to show the types of equilibrium that can occur when two sectors are completely autonomous, or when they are completely intertwined. An extension shows a possible pattern of regional growth in which one region harbors all the new firms, and older firms are relegated to another region.

In chapter 4, we look further at the theoretical predictions of the model in which firms are linked through an input-output matrix. In this chapter, we abandon the fluid sector-concept and look at the theoretical qualities of a model with discrete sectors. We present a method of determining the type of equilibrium that a model will attain and show how it depends on the value of the input-output matrix. Rather than presenting a few cases, we map the entire space of possible IO-matrices into the four types of equilibrium. The borders between these four types are such that 'dramatic' changes in equilibrium can occur. This means that, just as a small change in transport costs can precipitate a big change in equilibrium, so can a change in the IO-parameters.

Having explored the theoretical properties of the different models, we put them to an empirical test in chapter 5. We discuss the different methods that have been used to test models of economic geography in the literature. Using data on American states, we then parameterize a model of economic geography for the USA. We present two methods and use them on the same dataset. The first method mimics a study by Redding and Venables (2001), which uses a two-step procedure to assess the influence of economic geography-variables on regional wages. Where Redding and Venables have considerable success with data on different countries, we find that our analysis with data on American states is slightly less successful. Our second method obeys the general equilibrium conditions of the model, but is more computationally intensive. We use the parameters from this estimation to run a number of counterfactuals, showing the effect of infrastructural changes on different regions.

Chapter 6 reports the results of a policy evaluation exercise carried out in the summer of 2000, wherein the construction of a high-speed rail connection between the West and the North of the Netherlands was studied. The study uses the estimation methods from the previous chapter and combines them with detailed, regionally dependent IO matrices. We construct a model with considerable institutional detail in which different sectors and modes of transportation are identified. We are able to measure the direct and first-order indirect effects of the new railroad, but run into problems with the long-run solution of the model. An extension that models scarcity on the labor market is needed to come to a full solution.

Finally, chapter 7 summarizes the arguments and repeats the most important conclusions.

Chapter 2

A survey of complementarities in growth and location theories

2.1 Introduction

In our ever-changing economy, few trends last so long that they may be used to characterize the developments from the industrial revolution until today. Yet over the centuries, two phenomena seem to have stood the test of time: every year, on average, economic output grows by a few percentage points (Romer 1986). And, through the years, economic activity has always agglomerated into small areas, instead of spreading out evenly (Krugman 1991a).

As an example of continuing growth, consider figure 2.1 on page 10 which shows Dutch GDP per capita over more than a century. On average, growth is about 1.5% per year and apart from the period 1930–1945, the crisis and war years, economic growth is a regular phenomenon. In figure 2.2 on page 11, observe the year-2000 production per hectare[1] for each of the 12 Dutch provinces. There is an astounding eleven-fold difference between the most and least producing province. The differences in total production reflect different populations as well as differences in productivity, but are not necessarily the results of exogenous differences in natural endowments of the provinces.

This thesis is concerned with the things that economic science has to say about these two matters. What exactly brings about the growth of an economy, and why is it that production is so unevenly distributed? Using economic theory and models, we try to answer these questions and evaluate how policy affects growth and concentration.

In terms of models, the treatment of both growth and agglomeration has been rather upside-down. As for growth, the Solow (1956) model explains transitory adjustment processes, but the persistence of growth is an

---

[1] One hectare is 10,000 m^2 and approximately 2.47 acres.

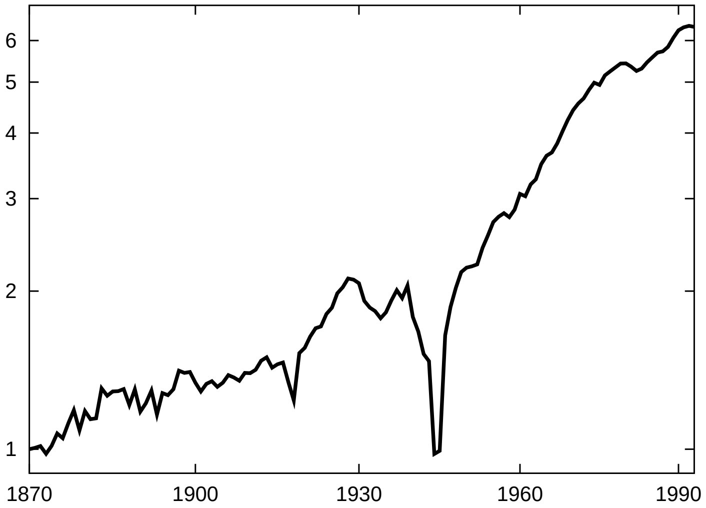

Figure 2.1: GDP per capita in the Netherlands, index numbers $(1870 = 1)$, log scale. Source: Maddison (1995)

assumption rather than an outcome. Agglomerations can be studied using land-rent models based on von Thünen (1842) or using the central-place theory of Lösch (1967). The former shows how the existence of a center affects the hinterland, while the latter constructs the efficient placement of centers on a featureless plain. In both theories however, the center is assumed rather than derived.

There exists an interesting connection between the deficiencies of these two approaches, and it is this connection that will be the theme of this survey. The inability of both models to generate the phenomena that seem so characteristic of real life is caused by the market form that is used. Both assume that economic activity is exclusively conducted by firms that are in full competition. This market form is in accordance with the firms' technical specification, namely, it is assumed that all firms are subject to constant returns to scale. In the context of both growth and location theory, we will contrast models which use CRS[2] with alternatives that feature increasing returns to scale and monopolistic competition. These models use the Dixit and Stiglitz (1977) MC framework to construct theories where growth and agglomeration are a consequence of the model, rather than an assumption.

---

[2] When no ambiguity may arise, I use the expressions 'constant returns to scale,' 'constant returns,' and the acronym CRS interchangeably. The same holds for MC as an acronym for Monopolistic Competition.

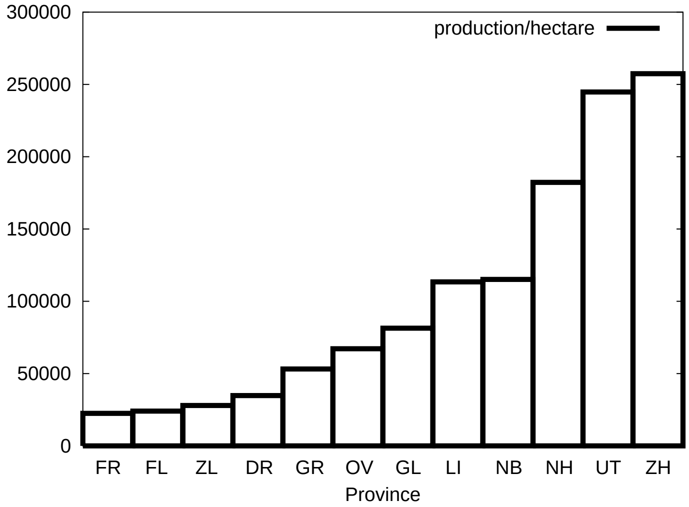

Figure 2.2: Gross regional product in Euros per hectare for twelve Dutch provinces, 2000. Source: CBS (2002)

Growth becomes endogenous growth, and the city center becomes an endogenous agglomeration.

The MC model has caused quite an upheaval in many areas of economics. The different fields that have profited from this innovation are discussed in Buchanan and Yoon (1994) and Brakman and Heijdra (2003), for instance.

As for the CRS assumption, it is easily seen that constant returns to scale severely limit the possible outcomes. If production is conducted under CRS, each separate factor in production faces decreasing returns. When growth is based on the accumulation of a subset of factors, this means that the economy cannot grow without bounds, and ends up in a steady state with zero growth.[3]

In location theory, the spatial impossibility theorem of Starrett (1978) (which can be found in Fujita 1986) states that a model with mobile agents on a closed, homogeneous space, facing a CRS production technology, can never explain the occurrence of agglomerations. Land rent will disperse economic activity without any countervailing force, because dividing up production over many locations leads to no loss in efficiency.[4]

Given these shortcomings of the CRS framework, it would seem tempting to use a wider class of firms, including those with increasing returns to

---

[3] For a detailed analysis, see section 2.4.

[4] The clash between economic geography models and the need to specify the market structure is discussed at length in Krugman (1995a).

scale. However, with the relinquishing of CRS, the assumption of full competition becomes untenable. The occurrence of other market forms greatly complicates the analysis, and allows for few analytical results. Fortunately, as we indicated above, a concise model of monopolistic competition introduced by Dixit and Stiglitz (1977) can be used to circumvent these problems.

This paper will briefly look at the monopolistic competition framework, and surveys the endogenous growth theory and economic geography in the light of it. It turns out that many interesting results in the two branches of literature can be attributed to the same fundamental properties of the monopolistic competition framework. The interplay between growth and geography is therefore not purely coincidental. While the models that show this were only recently made rigorous, their conclusions have been anticipated decades ago by such economists as Kaldor (1970) and Myrdal (1957):

> "[...] the movements of labour, capital, goods and services do not by themselves counteract the natural tendency to regional inequality. By themselves, migration, capital movements and trade are rather the media through which the cumulative process evolves—upwards in the lucky regions and downwards in the unlucky ones." Myrdal (1957, p. 27)

The antiquated notion of 'cumulative causation' is revived today as a process caused by complementarities in the model.

I will first look into the nature of monopolistic competition and the complementarities that characterize it. This is done in section 2.2. The findings are then used to provide a selective survey of economic geography (section 2.3) and endogenous growth theory (section 2.4). The two strands of literature are brought together in section 2.5, and section 2.6 concludes.

2.2 Complementarities and the monopolistic competition framework

In the classical framework of economics, many important results are obtained under a broad set of assumptions. For instance, the propositions of welfare economics as they may be found in Arrow and Hahn (1971) or in Takayama (1985, p. 185), guarantee that in general, decentralized market outcomes are socially optimal.

The theory assumes, among others, that all producers of goods are in full competition. This assumption implies a number of important simplifications: under full competition, one producer's pricing decision does not influence the market price. Also, no producer makes a profit, and prices should equal marginal and average costs. These simplifications produce, in general, a single, unique, welfare-maximizing solution. They also allow for simple pricing rules in the absence of strategic considerations. In this environment, a great number of analytical results may be derived.

As noted by Dixit and Stiglitz (1977, p. 297), the existence of a unique and optimal market equilibrium can challenged for at least three reasons, one of which is a failure of the model to reflect economies of scale that are observed on the level of a firm.[5] Allowing for these economies of scale, however, means letting go of the CRS assumption. This alters the behavioral assumptions that are appropriate for the firms (Helpman 1984). Increasing returns imply, for instance, that the largest firm has the lowest average costs, and is able to push the smaller competitors off the market. Even if this may seem realistic for some sectors, it makes it much harder to derive analytical results.

There is a case for abandoning CRS however. The assumption tends to bend reality, and paints a world in which economic transactions are basically a zero-sum game. In a CRS economy, it is of no consequence whether all people divide their time over the same range of activities, or whether each person specializes in a single activity and trades with the others. Clearly, this outcome is unsatisfactory as a reflection of real economic activity. It goes as much against common sense as it goes against the founding words of economics as a science, dedicated to the productivity gains from dividing labor (Smith 1776, p.13).

The issue whether to assume CRS thus turned out to be rather crucial for a coherent model of general equilibrium, but unrealistic in practice. This left economists divided for a long time:

> '... there seem to be two traditions, which persist. On the one hand there are those who are so impressed by what has been done by the CRS method that they have come to live with it; on the other, those for whom scale economies are so important that they cannot bring themselves to leave them aside.' (Hicks 1989, p. 12)

Among the efforts to bridge the gap was the work by Chamberlin (1933) and Robinson (1933), who sketched an alternative market form to the full competition implied by CRS. Their framework, monopolistic competition, held the promise of reconciling the two camps, but was rejected by most economists because of supposed inconsistency (Heijdra 1997). A severe blow to the MC model was dealt by Stigler (1968), who considered the model a failure and argued that economists should restrict themselves to the analysis of perfect competition and monopolies.

Despite these problems, the attractions of the MC model remained such that economists kept searching for a formulation that would be both math-

---

[5] The other two are distributive justice and external effects.

ematically consistent and useful in practice. This would be a version where strategic interactions between the different firms would have to be somehow 'sanitized' from the model, for it was clear that such considerations would severely cloud the search for equilibrium. A formulation that allowed exactly that was finally found by Spence (1976) and Dixit and Stiglitz (1977). Their breakthrough articles set a chain of events in motion in which the MC-alternative to CRS became widely used, especially in industrial economics, trade theory, growth theory and economic geography. Although theirs is 'a very restrictive, indeed in some respects, a silly model' (Krugman 1998, p. 164), allowing the economist to focus on the effects of increasing returns without worrying about strategic interactions between firms made it an instant classic. The apparent arbitrariness of the model is not denied, but taken for granted, hoping that insights will extend beyond the model:

> "Unfortunately, there are no general or even plausible tractable models of imperfect competition. The tractable models always involve some set of arbitrary assumptions about tastes, technology, behavior, or all three. This means that [...] one must have the courage to be silly, writing down models that are implausible in the details in order to arrive at convincing higher-level insights." (Krugman 1995a, pp. 14-15)

It is important to realize that the monopolistic-competition approach is not the only available route into increasing returns, and that some insights are sacrificed when it is chosen. As Dixit and Norman (1980) write,

> For descriptive purposes, one must [...] choose among the numerous alternative ways in which imperfect competition can be modelled; and the conclusions one arrives at will in general depend on the particular specification chosen. [...] The best one can hope for is a catalogue of special models. (p. 265)

Neary (2003) argues that the MC model has nothing to say, for instance, about the effects of globalization on market structure. In that case, a model of strategic oligopolistic interaction is needed.

This section provides a short introduction to the Dixit-Stiglitz monopolistic competition framework. Before looking at the model itself, we will briefly discuss the problems that surround returns to scale in general, and the notion of externalities.

2.2.1 Returns to scale

A firm's production possibilities are summarized in its production function. If for an amount A of a certain product a firm uses inputs, whose quantities are summarized in a vector $\mathbf{B}$, the correspondence between different values of A and $\mathbf{B}$ defines the production function $f(\mathbf{B})$. For any $\mathbf{B}$, we can evaluate the returns to scale of the firm by looking at the point elasticity

$$\varepsilon_B = \left.\frac{\partial f(\lambda\mathbf{B})}{\partial\lambda}\frac{\lambda}{f(\mathbf{B})}\right|_{\lambda=1}.$$

When ε_B is larger than one, there are increasing returns to scale. Note that ε_B is a function of the inputs $\mathbf{B}$. A firm can have increasing returns for all possible $\mathbf{B}$, but also for a limited set of values of $\mathbf{B}$.

On the level of the entire economy, increasing returns to scale are fairly undisputed. In this case, we can think of f as a nation's production function, with $\mathbf{B}$ indicating the supply of labor and capital. Increasing returns have been attributed to the division of labor (Smith 1776), splitting up complex production methods into multiple simple steps (Young 1928, Stigler 1951), and the fact that technological knowledge, once produced, is nonrival and nonexcludable (Romer 1990). It would be a positive quality of any economic model to have the possibility of including increasing returns on the macro level.

Much of today's macroeconomic theory is derived explicitly from microeconomic foundations (see, for instance, Romer 1993). The occurrence of increasing returns at the micro-level spells trouble. Helpman (1984) shows that the modeler needs to specify a host of parameters to even start working: the conditions of firm entry, the heterogeneity of the good, and the type of market are just a few among them. The outcome of the model is highly dependent on these assumptions, for instance, do firms compete in a Bertrand– or a Cournot–market?

The simplest of these assumptions is that every sector is dominated by a single monopolist, who fully exploits the increasing returns. Apart from the question of realism, the presence of monopolists causes problems in a general-equilibrium model. One source of problems is the occurrence of monopoly rents: the model needs to specify how these rents are spent by the monopolist. In full competition, profits are zero by definition.

To avoid these issues altogether, one can assume that part of the returns to scale are external to the firm. The idea, originally from Marshall (1920), separates internal economies ('those dependent on the resources of the individual houses of business engaged in it', p. 266) from external economies ('those dependent on the general development of the industry', p. 266). External economies, or externalities, do not affect the firm's optimization; thus, they can be incorporated in a consistent profit-maximizing framework, where firms perceive their situation as one of full competition. Between externalities, we can find two types (Scitovsky 1954): pecuniary externalities, those which are mediated by markets, and the rest, non-pecuniary externalities.

Non-pecuniary externalities use a production function, at the firm level, like $f(\mathbf{B}) = \tilde{f}(\mathbf{B}, X)$. Here, $\mathbf{B}$ again are the inputs and X is industry output (Helpman 1984). Every single producer considers X as given, and controls only $\mathbf{B}$. But f may have increasing returns in $\mathbf{B}$ and X together.

Using non-pecuniary externalities, it is possible to construct a model of general equilibrium that features increasing returns. Although this has indeed been done (Chipman 1970), such models have not been used extensively. By their nature, non-pecuniary externalities are not observed so that the economist can assume anything about them. Any possible outcome can thus be 'doctored' into the model.

Pecuniary externalities are more subtle. It could be possible that a producer, by entering a market, increases the consumers' utility because of the increased variety that he/she provides. Although profit opportunities were the firm's original motive for entering, the variety effect may influence the perceived price level faced by the consumer, and alter the allocation of goods. Another example would be the entry of a firm that, because of its demand for an input, affects the price that input for all other firms.

However, the methodological problems outlined by Helpman (1984) still need to be solved. A particular model that knits together increasing returns at the firm and macro level in a consistent way, and thus solves these problems, is the Monopolistic Competition model of Dixit and Stiglitz (1977). The introduction of this model, in which pecuniary externalities drive the equilibrium, for the first time allowed the analysis of increasing returns and caused what Brakman and Heijdra (2003) call the 'second[6] monopolistic revolution.' We will introduce the MC model in the following section.

2.2.2 Monopolistic competition

The key difference between full competition and monopolistic competition[7] is in the nature of the traded good. With full competition, the good is assumed to be homogeneous, and its price the only criterion of selection. With MC, consumers discern different varieties, and products from different producers are imperfect substitutes.[8] Even if each individual producer faces increasing returns to scale in production, the largest producer is not always able to push smaller competitors out of the markets because substitution between products is limited.

In most applications of MC, consumer preferences are modelled as in

---

[6] The first monopolistic revolution was the idea of MC being formulated by Chamberlin (1956).

[7] We will use the acronym MC for 'monopolistic competition' from now on.

[8] Chamberlin (1956, p. 56) suggests that such elements as 'the conditions surrounding its sale', trade marks and the seller's reputation 'may be regarded as [being purchased] along with the commodity itself.'

Dixit and Stiglitz (1977)[9]. The quantities of goods x_i consumed are aggregated in a CES function,

$$U\left(x_1,\ldots,x_n\right)=\left(\sum_{i=1}^{n}x_i^{\theta}\right)^{1/\theta}. \tag{2.1}$$

with $0<\theta<1$. By choosing suitable units of measurement for the different goods, we can abstain from adding scale parameters to the different x_i. It is clear that for each of the goods, an increase in the amount consumed will increase total utility. If we maximize (2.1) with respect to a budget constraint $\sum x_i p_i = E$, we find that

$$x_i=\frac{E}{q}\left(\frac{p_i}{q}\right)^{-\sigma} \tag{2.2}$$

where $\sigma = 1/\left(1-\theta\right) > 1$, and we have used the associated (ideal) price index $q=\left(\sum p_j^{1-\sigma}\right)^{1/(1-\sigma)}$. We assume a large number of producers n so that the effect that one producer's price has on q is vanishingly small. So, each producer takes the price index as given and faces a demand elasticity σ for his product. Also, he does not need to take the behavior of other producers into account when deciding on price and quantity. Strategic motives are absent, and this makes the model tractable and easy to solve.

If every variety sells for the same price p, all are purchased in the same amount. In this case, formula (2.1) shows that utility is $n^{1/(\sigma-1)}E/p$. That is, an increase in variety brings an increase in utility even if the nominal budget remains the same. Helpman and Krugman (1985, p. 117) call this the 'love-of-variety effect.'

The more varieties (n) there are, the less influence a single producer's price exerts on the consumer's real income. To completely eliminate every producer's market power, it is often assumed that the range of goods $[0\ldots n]$ is continuous, and each producer is infinitely small. Though awkward, this assumption can be given some rigor. This is done in appendix 2.A.

Producers are usually assumed to face a fixed cost for setting up production and a variable cost per item produced. This implies the average cost per product declines with total production, so that producers are subject to increasing returns technology. This encourages firms to expand their output as much as possible; however, they also face a downward sloping demand curve as we saw in formula (2.2). Thus, producers maximize profits by setting marginal benefit equal to marginal costs, which given (2.2)

---

[9] Weitzman (1994) shows that this model is much related to the Lancaster (1979) 'spatial competition' model, where each consumer has an ideal product and picks the one closest to it.

results in a mark-up over marginal costs of size $1/\theta$. In equilibrium, all producers set the same price. The number of active producers adjusts so that discounted profits are just enough to recoup the initial investment F. With free entry, this means that n adjusts to drive profits to zero.

The constant elasticity of demand, faced by a producer, is at once an advantage and a disadvantage of the model (Dixit 2000). It allows us to get a simple form for the pricing equation, which gives the model much of its appeal. However, as the number of varieties increases, we would expect the products to become more similar and the elasticity of demand to increase. This way, there would be a competitive limit to the model. In the current formulation, this is not the case. We should recognize this flaw when we discuss models where n grows *ad infinitum*.[10]

In an alternative interpretation of the same model, Ethier (1982) used the aggregator function in (2.1) as a production function. Output U is made with inputs x_i; each input is produced by a single intermediate goods producer. The production function belongs to a class of firms that convert the intermediate goods into a final consumer good. These firms face constant returns to scale, as may be checked from (2.1), and are in full competition. The 'love-of-variety effect' from above has now become quite another thing: when entrance is free, there are increasing returns to scale at the economy's macro level. We will return to this interpretation below, as well as in the following chapters.

Now that increasing returns to can be modelled consistently, we are able to construct a general equilibrium theory where the actions of one firm affect the conditions of other firms, though not intentionally. We will find that many equilibria in MC models, for their stability, depend on the fact that the actions of several firms complement each other. Complementarity is the subject of the next section.

2.2.3 Complementarities

Matsuyama (1993, 1995) discusses complementarities, the notion that "two phenomena (or two actions, two activities) reinforce each other." (1995, p. 702). Complementarities often arise in the MC framework.

As a specific example, assume that in an economy, people consume a single final product that is made out of several intermediate goods with production function (2.1). That is, there are n different intermediate goods, and total production is U. This is the Ethier-setup from above. Assume also that intermediate-goods producers face fixed costs F and variable costs θx_i

---

[10] An extension of the model that goes into this direction is introduced by Heijdra and Yang (1993)

which are both incurred in labor, that is,

$$\begin{aligned} L_i &= F + \theta x_i && (2.3) \\ \pi_i &= p_i x_i - w(F + \theta x_i) && (2.4) \end{aligned}$$

where x_i is the output of firm i (the double use of parameter θ is here and in formula (2.1) is for mathematical convenience), π_i is firm i's profit and w is the wage rate. Remembering that price, in this model, is a markup $1/\theta$ over marginal costs, we can use the elasticity of substitution in the price equation, writing it as

$$p_i = \frac{\sigma}{\sigma - 1} \cdot \theta w = w \tag{2.5}$$

where σ is defined as above. From this and (2.4) it follows that a firm that makes zero profits employs $L_i^* = \sigma F$ workers. When there are L workers in the economy and there is free entry in the intermediate sector, it follows that the number of producers in that sector will be

$$n^* = \frac{L}{\sigma F} \tag{2.6}$$

The production of the final good, per capita, is increasing in n^*, because of increasing returns to scale on the macro level. In fact, per capita production is $(n^*)^{1/(\sigma-1)}$.

Now if there exist two of these economies, with different intermediate goods, and they open up for trade, both economies will see the range of available intermediate goods increase. Because of this, both economies will experience an increase in production per capita. When the two economies interact, they are complementary to each other. This principle has been the basis for a large class of trade models, for instance in Helpman and Krugman (1985).

Hirschman (1958) discusses a related issue in the context of economic development. In his terminology, there exist *linkages* between different firms in a region. These linkages concern the input-output relations among the firms. Hirschman distinguishes *backward* linkages when a firm demands inputs from other firms, and *forward* linkages when a firm produces inputs for other firms. The conjecture is that with positive costs of transport for intermediate goods, linkages between firms can make an agglomeration stable.

In fact, the conjecture requires that linked firms are complementary to each other. It is true that in general, the arrival of a downstream firm can induce an upstream firm to expand. However, when this happens in a constant-returns world, the expansion has no effects on the original activities of the individual upstream firm, and merely leads to entry of upstream firms. The linkage is rather weak in this case. But should the upstream firm exhibit increasing returns to scale, expansion means that it can now operate at a higher level of efficiency. In that case, the two firms are complementary.

2.2.4 Review, and a look ahead

To study a complex phenomenon, it can be necessary to make a number of assumptions that simplify the problem. We have argued that the CRS assumption fulfilled such a role in economics, as it allowed the derivation of a simple rule of conduct for firms, namely, marginal cost pricing. It also solved the problem of which market form would prevail, in favor of full competition.

We have also introduced an alternative framework, based on a different assumption: the MC setup. This setup is not any more general than full competition, the number of assumptions has even increased. Yet it is an interesting alternative because it allows for complementarities and increasing returns to scale.

The short introduction above does not do justice to all the intricacies of MC, but that is not the point of this survey. Rather, we now want to look at the application of this framework to two fields, economic geography and growth theory. The application of MC to these fields has allowed a large number of innovations. Those in economic geography are discussed in the following section, while those in growth theory are the subject of section 2.4. The two strands of literature are brought together in section 2.5.

2.3 Economic geography

Ironically, economic geography or location theory has been a rather peripheral field of study within economics. In part, the small amount of attention for issues of location can be attributed to the institutional, geographical and sociological factors that play such an important role in the problem. Yet over the years, many interesting results have been obtained using methods of economics. We look at the foundations of location theory in section 2.3.1. Then we turn to a new class of models that involve monopolistic competition and increasing returns in section 2.3.2.

2.3.1 Foundations of location theory (1): Exogenous agglomeration

The earliest theory of location can be divided in two branches (Greenhut 1956): least-cost theory, oriented on the supply side, and spatial competition theory, oriented on the demand side of the economy. The striking characteristic of least-cost theories is that they start by assuming a form of agglomeration; they do not explain why the agglomeration came about in the first place. This problem is tackled to some extent by spatial competition theories, as well as by the theories based on externalities and those that use increasing returns. We look at the different theories in chronological order.

Least-cost or land use theory starts with assuming that all demand in the economy is located at a single point. This can be a mining town demanding agricultural produce as in von Thünen (1842), or a central business district in which all trade is conducted, as in Fujita (1986). Transportation is costly, and costs increase with distance from the center, r. From their production function and the costs of transport, suppliers can compute how much rent they want to pay as a function of r. This information is aggregated in a rent gradient, according to which the suppliers settle. The approach is refined by Weber (1909) to account for the location of raw materials, and Alonso (1964) adds, among other, endogenous lot size. Many models of urban structure still use this setup.

Spatial competition or locational interdependence theory, on the other hand, does not assume the existence of a center. Rather, (consumer) demand is distributed over locations and (zero-size) producers are looking for the optimal spot. With land rent out of the model, this approach clearly deals with questions of attraction and repulsion among different firms, which places it in the realm of game theory. The founding paper of this field is Hotelling (1929), who shows that two producers of a homogeneous good will locate next to each other halfway a line with evenly spread consumers. This is not the socially optimal situation. Chamberlin (1956, pp. 260-265) shows that increasing the number of sellers in this problem will cause their dispersion, converging to the optimal dispersion as the number of sellers goes to infinity. Gabszewicz and Thisse (1986) provide a survey of this method.

The two approaches above may be combined. Lösch (1967) and Greenhut (1952) introduce profit-maximization as the relevant criterion. Given that demand and supply conditions may vary with location, this tends to make the problem less tractable. There exist fewer general rules on spatial dispersion than in the above, simplified, analysis. An important limitation of both these approaches is the assumption that consumers do not change their location in response to the suppliers' whereabouts.

2.3.2 Foundations of location theory (2): Endogenous agglomeration

The location theory in the preceding section has said very little about the causes of agglomeration. We can think of at least three types of forces that drive people and firms to the same location. Firstly, there are the autonomous characteristics of the landscape. Some places may be more pleasant as a place of residence, or productive as a place of business than others. Natural harbors or strategic points fall in this category.

Secondly, it is often thought that nonmarket externalities are an important factor in the creation of agglomerations. Such hard-to-measure concepts as informational and technical spillovers between firms, or in general informational exchanges between agents (Fujita and Thisse 1996, p.347) cause people to cluster together. The reason for clustering is the fact that the amount of spillovers between two firms is assumed to decline rapidly with distance. The spillovers are embodied in such acts as face-to-face talks and casual inspection of the other firm's production site. Nonmarket externalities are emphasized in Jacobs (1969).

The problem with the above two conjectures about the causes of agglomeration is that they are difficult to verify. Saying that agglomerations are caused by agglomeration economies is close to a tautology. Designating the spots where people have clustered 'attractive' is not much better. The predictive power of these theories is small. It is therefore preferable to have a model where agglomerations are a result of more fundamental properties like the way people consume and produce.

Recently, economists working on the theory of (urban) agglomeration have put forward a number of such micro-foundations. Duranton and Puga (2003) classify them using three motives for agglomeration: *sharing*, *matching* and *learning*. In each of these classes, models are formulated featuring pecuniary externalities. In the class of models in which *sharing* is the driving force behind agglomeration, the monopolistic competition framework of section 2.2.2 is the preferred vehicle of analysis. We briefly discuss each class of motives here, before concentrating on a particular kind.

- **Sharing**. The agglomeration of a large number of people into a city may be explained by the presence of an indivisible service to consumers such as a stadium, or by production-side indivisibilities such as a large factory. However, even with smaller sized services and firms, the *variety* they constitute when gathered into one place may be a force of attraction. Firms are complementary to each other in this respect, each one constituting a small part of the total supply of variety. We will spend most of this section on a model that shows how such sharing of variety may lead to equilibrium agglomeration. Other models in the *sharing* category use the returns to specialization and the sharing of (labor market) risk, a subject that goes back to Marshall (1920).

- **Matching**. Another way to bring out the advantages of a large labor pool is by looking at the heterogeneity of labor. Both the probability and the quality of labor market matches are larger when the number of firms and laborers is big. Thick labor markets in cities can thus play a role in the stability of the agglomeration.

- **Learning**. Finally, the role of cities as repositories of knowledge can be used to explain their success. Non-pecuniary effects in knowledge generation as emphasized by Jacobs (1969) fall into this category, as do models of skill transmission. Finally, models of learning-by-doing

> with specific urban features exist. These will be discussed in section 2.5.

In the rest of this section, we will concentrate on a model in which the variety offered by different producers leads to agglomeration. Because of their differences, producers are complementary to each other and when different producers are located at the same spot, their combined presence is an equilibrium. The principle, complementarity-induced agglomeration, was recognized by Krugman (1979) in a paper about monopolistic competition and international trade. In his model, where the MC setup was slightly different from above[11], when trade was prohibited but factors were mobile,

> [...] there will be an incentive for workers to move to the region that already has the larger labor force. [...] In equilibrium all workers will have concentrated in one region or another. (p. 20)

It is not very difficult to see the agglomerative tendencies using the model from section 2.2.2. Suppose that there are two regions in which economies with an MC structure exist, and that trade is prohibited. The number of firms in each region is linear in the number of inhabitants (section 2.2.3); the aggregate price index faced by each inhabitant is (section 2.2.2)

$$q = \left(\sum_{j=1}^{n} p_j^{1-\sigma} \right)^{1/(1-\sigma)} = \left(n p^{1-\sigma} \right)^{1/(1-\sigma)}$$

which is decreasing in n (this is due to the 'love-of-variety-effect'). If inhabitants are given the choice where to live, they will move to the more populated region. Hence, agglomeration results naturally.

The first to design explicit models of location based on MC were Fujita (1988) and Rivera-Batiz (1988). These models featured agglomeration economies as well as land rents based on a least-cost framework (see Section 2.3.1 above). Despite the countervailing force of the rents, Rivera-Batiz shows that for some parameters, "[t]he economy's population [...] ends up completely in city m" (1988, p. 148).

Since so much of this book will be based on this type of models, we take the time to specify an economic geography model in the next section, and characterize the different varieties that are known.

---

[11] Specifically, the subutility-function x^{θ} is replaced by a function $v(x)$. The elasticity of demand is now $-v'/v''x$, and it is assumed that the elasticity decreases in x (this does not happen with the x^{θ} form). The assumption leads to the result that wages are higher in the most populated region; with the x^{θ} form this is not true, unless the wages are corrected for the local price index.

2.3.3 The Core-Periphery model and the home market effect

Endogenous agglomeration is the hallmark property of class of models known as 'new economic geography.' The complementarities that cause agglomeration arise because of the MC framework, but they may travel through different markets. In this section, we will look at the model of Krugman (1991b) where complementarities go through the labor market. We follow the exposition of Neary (2001), which is a little more concise than the original.

There are two types of production, agricultural and industrial. The former is homogeneous and operates under CRS, the latter is MC. Consumers in either region maximize utility,

$$\bar{U} = A^{1-\mu} \cdot U^{\mu} \tag{2.7}$$

where A is consumption of agricultural products and U is the aggregate utility obtained from the consumption of different varieties of the industrial product. U is defined in equation (2.1) above. We deduce that μ is the share of income spent on manufactures. Because of the MC assumption, producers face a demand schedule as in (2.2), where they take the price index q as given. The budget E is the amount of money that consumers spend on industrial goods, or μ times their income. This form of the demand function leads to mill pricing, where the price is a markup over marginal costs.

The production function uses only labor and is given in (2.3). This leads to the result that production per firm is $y_i = \sigma F$, and thus does not vary with any endogenous variable. This result is rather remarkable, and rather special. It requires that increases in costs because of rising wages, for instance, are exactly balanced against increases in revenue because of rising prices. A number of these special properties makes the MC model tractable, but very special. Because of the fixed output per firm, shifts in the level of total production are caused by changes in the number of firms, n^*, which is defined in (2.6).

Up until now, our model is a standard full equilibrium MC model of a closed economy. This changes when we assume that there are two regions for which the above specification holds, and that there are positive transport costs between them. These costs of transport take the *iceberg* form, where they are incurred in the shipped product itself. We assume that of what is shipped between the two regions, only a fraction τ actually arrives. The analogy of this type of transport costs to a melting iceberg is due to Paul Samuelson.[12] Agricultural goods are shipped without incurring transport costs and serve as the numéraire.[13]

---

[12] A similar analogy recently occured at a diner with a number of economists, one of whom refused to pass a bottle of wine without pouring some in his own glass.

[13] The assumption about the absence of transport costs for agricultural goods is not in-

The demand for industrial goods from the other region is different from the home demand (formula 2.2) for two reasons. Firstly, the price that is faced by foreign consumers is higher because of transport costs. Secondly, for each unit that is to be received in the other region, $1/\tau$ units must be shipped. This leads to the following form for foreign demand:

$$\begin{aligned} x_i^* &= \frac{E^*}{q^*}\left(\frac{p_i/\tau}{q^*}\right)^{-\sigma}\frac{1}{\tau} \\ &= E^*(q^*)^{\sigma-1}p_i^{-\sigma}\tau^{\sigma-1} \qquad (2.8) \end{aligned}$$

In these equations, foreign variables are indicated with an asterisk. From this, we can immediately see the use of the iceberg-assumption. Even though the foreign price is increased by transport costs, the elasticity of foreign demand with respect to p_i remains equal to σ. Thus, the maximization problem for the producer remains the same, as does the optimal price.

Of course, the price indices q and q^* change when trade between the regions is allowed. The home price index now takes the form

$$q = \left[np^{1-\sigma} + n^*\left(\frac{p^*}{\tau}\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}} \qquad (2.9)$$

We can write a similar expression for q^*.

Notice the mathematical form of (2.9), which has a sum inside a power expression. This form cannot be simplified through manipulation, and remains at the heart of many variants of the Core-Periphery model. This makes it impossible, in general, to solve the model analytically, which leads to the fact that many results have to be derived through numerical simulation. Krugman (1998) calls the need for reliance on numerical results one of the hallmark properties of economic geography models.

We pause for a moment to summarize what we have constructed thus far. We have two regions which produce agricultural and industrial products, both of which can be traded. Prices are set the same as in autarky, courtesy of our special assumption about the form of transport costs. The number of farms and industrial producers is governed by the zero-profit assumption. We have not made any special assumptions yet about the labor force and whether it can change sectors, but so far we have assumed that there is no interregional migration.

Consider now a situation where both regions are exactly equal. What will happen when the home region faces an increase in demand? We know that, since per-firm production and prices are fixed at their optimal level, the demand shock leads to the entry of new industrial firms. These firms

---

nocuous. Davis (1998) shows that the introduction of positive costs of transport may negate several results, including the Home Market effect (see below and appendix 2.B).

produce new varieties of the industrial product, which will lead to a change in home's price index q. This in itself leads to a change in the equilibrium, which could result in a further bout of entry or exit. Will the demand shock lead to a more or less than proportional increase in the number of firms?

In appendix 2.B, we derive a result known as the *home market effect*, which answers this question. The result, originally by Krugman (1980), says that the region with higher demand will have a disproportionally *higher* number of firms. That is, the change in q leads to the entry of even more firms. From this, we can conclude that regions with a larger home market will be net exporters of manufactures. This result has been the basis of empirical tests of the MC model, which are discussed in section 5.2.1 below.

The home market effect tells us that an increase in demand leads to an even larger increase in the number of firms. If it were true that an increase in the number of firms, through some channel, caused another increase in demand, we would have a closed causal loop that could explain spontaneous agglomeration and persistence of regional differences as discussed in the quote by Myrdal on page 12.

There are a number of different possible channels through which the causal loop can be closed, and the choice of channel defines the type of CP model. Ottaviano and Puga (1997) classify three different types of models according to these media. They discern migration linkages, input-output (or intermediate good-) linkages and intertemporal linkages. We discuss the three types of models in turn. In each of them, the basic setup is as described above: there are two regions and each region has two sectors, agriculture and manufacturing, which are competitive and MC, respectively.

2.3.4 Three channels in the Core-Periphery model

The migration-based CP model

The first CP model, constructed by Krugman (1991a, 1991b), is based on migration linkages. In it, manufacturing workers decide where they want to live based on their real wage. From (2.7) and the fact that the price of agricultural goods is normalized at one, we find that

$$\omega = \frac{w}{q^{\mu}} \tag{2.10}$$

for real wage ω.

Agricultural workers are supposed to stay put. To see how this changes the model, remember the hypothetical shock in demand of the previous paragraph: it led to a (disproportionally large) change in the number of firms, but there the causality stopped. Now, the changing number of firms will also, because of the variety effect, alter the price index of industrial goods. This in turn affects real wages and will lead to migration. And the channel may not end there: we will have to trace the effects of this migration, to find out where, or whether, the process ends.

Neary (2001, p. 542) finds that there are three effects of a change in the number of firms in a region. The first is the competition effect: when there are more firms, each gets a smaller piece of total revenue. This effect does not depend on labor mobility and was already present in the model above. It is a stabilizing force, in the sense that it limits the number of firms that can profitably enter after a shock in demand.

The second and third effect work through the mobile labor force. As the number of firms in a region increases, the increased scarcity of labor will drive up real wages, inducing migration into the region. This has two effects. Firstly, the new workers will demand manufactures from local producers. This *demand linkage* leads to an increase in profitability. Neary (2001) shows that, assuming wages return to their prior levels, the balance of the first and the second effect depends on the relative sizes of μ, the share of manufactures in demand, and Z, the index of transport costs defined in appendix 2.B. If μ is larger, the (destabilizing) demand linkage dominates, while a larger Z implies that the (stabilizing) competition effect is stronger.[14]

However, we must also take into account that the decline in price index q caused by the increasing number of firms, leads to a lower cost of living in the region where the demand shock took place. Since real wage ω must be equal in both regions, the assumption that wages return to their previous level must be false. Nominal wages can fall, leading to a further increase in profitability. This is the third effect.

The balance of (stabilizing) effect 1 versus (destabilizing) effects 2 and 3 determines whether a symmetric equilibrium, in which both regions have the same number of firms, can be stable. If a small demand shock in one region leads to a cumulative process of migration and firm entrance, the equilibrium is unstable; if instead it fails to lead to a cumulative process the equilibrium is stable. Using the properties of the model, we can derive a condition on the parameters that tells us whether the symmetric equilibrium is stable or not. The level of transport costs at which stability changes is called the *break point* τ^B.

Similarly, we can ask whether complete agglomeration is stable. That is, when all manufacturing is concentrated in one region, and there is a demand shock in the 'empty' region, does a cumulative process ensue which leads to a symmetric equilibrium? If not, the agglomeration is stable and

---

[14] This result is appealing: a large Z corresponds to low costs of trade and diminishes the market power that producers exercise over local demand. With low costs of transport these consumers can easily substitute imported goods. A large value of μ indicates that consumers spend a big share of their income on manufactured goods, making their arrival more interesting to producers, but only to the extent that buying local goods is attractive.

the model returns to its previous state after the shock. There exist values of τ for which this is indeed the case, so that the model can explain endogenous agglomeration, as promised. However, for transport costs that are too high the equilibrium is not stable. Hence, a level of transport costs may be derived at which the stability of the concentrated equilibrium changes. This level is called the *sustain point* τ^S.

Both points are derived in appendix 2.C. It turns out that the sustain point and the break point are generally not the same, and that $\tau^B > \tau^S$. This means that there exist transport costs τ^+, with $\tau^B > \tau^+ > \tau^S$, where both the agglomerated and the symmetric equilibrium are stable. Which equilibrium actually occurs depends on the initial conditions: if the model starts off close to symmetric, the symmetric equilibrium will be attained. If the model starts out with all industry concentrated in one region, it stays that way. The CP model with transport costs τ^+ has a path-dependent solution.

The intermediate goods-based CP model

Venables (1996a), in a model where labor is not mobile, shows that it is possible that input-output linkages between firms fulfill the same role as a mobile workforce. Using a monopolistic competition setup for both an upstream and a downstream sector, Venables shows that it is possible that an increase in the size of one industry brings the other industry to a higher level of efficiency. The model's conclusions remain the same in Krugman and Venables (1995), who extend the framework by collapsing the upstream and downstream industries into one layer. The monopolistic competitive market structure is preserved by a specific form of the final demand function. Amiti (1997) shows that a similar outcome may be obtained without the use of an MC framework. In her model, a scale effect arises because of a pricing game that is played between firms in a sector. An increase in the number of firms has a negative effect on collusion and ups the sector's efficiency.

Later in this book, we will make good use of the model where intermediate goods transmit the complementarity between firms. A detailed introduction to this model can be found in chapters 3 and 4.

Intertemporal linkages and the CP model

Aspects of factor accumulation can also serve as a medium for agglomerative tendencies. Baldwin and Martin (2003) survey the interdependencies between agglomeration and growth; they divide the subject into two classes: in the first class, growth influences agglomeration but there is no causality going the other way. We will discuss this class in the current paragraph, as it illustrates how the accumulation of capital can lead to agglomeration. The other class, in which technological spillovers are only local, will be discussed after our introduction to growth theory below. In models of the second class, agglomeration can affect the rate of growth, and vice versa. They are the subject of section 2.5 on dynamic economic geography.

Assume that there exist knowledge spillovers, and that they are global. That is, the cost of capital investment declines as a function of the world stock of capital. In that case,

$$\dot{K}_i = \gamma \cdot L_i^I \cdot K_{\text{world}} \tag{2.11}$$

where the growth of the capital (or 'knowledge') stock of region i depends on the number of people working in the innovation sector, L^I and the world stock of capital K_{world}.

Whether capital accumulation of this kind can lead to full agglomeration depends on the mobility of capital. If we assume that inhabitants of one region can own and operate capital in another region, while spending the proceeds at home, capital is mobile. If instead we assume that most capital takes the form of human knowledge, which cannot be separated from its owner, capital is immobile.

Baldwin and Martin (2003) show that with perfect mobility, the initial distribution of firms and capital between regions is stable. Both regions save and accumulate capital, deploying it where it is most productive. With zero capital mobility, however, agglomeration in one region can occur. This happens when trade costs are sufficiently low.

The reasoning behind agglomeration is the following: agents can only invest in capital that is used in their own region. The incentive to invest depends on the profitability of operating a firm; the firm's profitability in turn depends on the demand for its products. Now if trade costs are high, local demand can be enough to sustain firms in either region. But with low trade costs, it is possible that one region enters into a downward spiral: if the number of firms declines, the income from capital declines (all capital is owned locally due to the immobility) which drives down local demand. Meanwhile, imports from the other region substitute for products that are no longer available locally. This further decreases the incentive to invest in local capital. Ultimately then, all investments are made in the other region.

2.3.5 Conclusion

Even though the mechanics, as well as the economic rationale of the above models are substantially different, there are some common characteristics that are worth spelling out. The most important outcome is that in all three models, the combination increasing returns - transport costs can in principle lead to agglomeration. Both are a necessary factor. If there are no increasing returns, firms may as well split up and be spread out over space without any loss in efficiency. If transport costs are zero, then the whole concept of location does not matter in economic decisions (this is the spatial impossibility theorem referred to in Section 2.1).

The relation between transport costs and agglomeration tendencies is often found to be an 'inverted U' (Junius 1996, Ottaviano and Puga 1997, Venables 1996a). At very high transport costs each region is self-sufficient and no interaction takes place. At intermediate transport costs the above agglomeration effects are stronger, and at very low transport costs the 'centrifugal' forces congestion and factor market competition take over and firms spread out again.

A natural question that arises when these theoretical models are presented is whether there is any empirical relevance to them. We will discuss the empirical literature that complements this theory in chapter 5, which deals with estimation. The survey is in section 5.2.

2.4 Endogenous growth theory

In the introduction, we spoke briefly about the inability of traditional theories of growth to explain lasting growth as an economic phenomenon. I will now substantiate these claims and introduce several alternatives that fall under the header of 'new' endogenous growth theory.

The MC framework that was introduced above does not play a pivotal role throughout endogenous growth theory. The new growth models were erected for a number of reasons, summarized by Romer (1994). Besides dissatisfaction with the inability of classical models to explain lasting growth, Romer identifies two other causes. One is the so-called convergence controversy: the (perceived) neoclassical prediction that poor countries must catch up with rich countries was disputed by data that became available around that time (Maddison 1982, Summers and Heston 1988). The other cause is the fact that the neoclassical model is at odds with a number of easily observable facts, facts which can only be explained if imperfect competition is incorporated.[15] As the MC framework was the first to allow imperfect competition to be modelled in a concise way, it has been the framework of choice for a lot of endogenous growth models.

There are basically two 'waves' of models within the theory; the first wave (started by Romer 1986) describes growth as a process of ceaseless

---

[15] The facts are: 1. There are many firms in the economy, not one monopolist. 2. Discoveries are nonrival. This makes them different from other inputs. 3. Physical activities can be replicated; therefore production functions should be homogeneous of degree one. 4. Technological advance comes from things that people do. It does not occur by itself. 5. Many individuals have market power and earn monopoly rent on discoveries even though they are nonrival: informaton can be excludable.

Classical growth models are at odds with facts 4 and 5. Not all endogenous growth theories accomodate all these facts.

accumulation of factors. It is possible to retain the assumption of perfect competition in these models, using externalities. We will look at a sample model that employs the MC framework, though. The second wave (started by Romer 1990) explains growth by organized technological progress, and uses the MC framework together with an explicit sector for R&D.

We first briefly look at the exogenous (Solow-) growth model and compare it with some first-wave endogenous growth models. We then look at the second-wave models in Section 2.4.2.

2.4.1 Neoclassical and endogenous models of accumulation

The macro level

The neoclassical growth model was developed independently in Solow (1956) and Swan (1956), and the setup can be summarized quite concisely. The economy of a country uses two factors, L and K, and produces a single output. A proportion $(1-s)$ is consumed[16], the rest is used to increase K:

$$\begin{aligned} Y_t &= F\left(K_t, L_t\right) && (2.12) \\ \dot{K}_t &= sY_t && (2.13) \end{aligned}$$

The aggregate production function F exhibits constant returns to scale, and the population of laborers L grows exponentially at rate n. To each factor L and K taken alone, the function has decreasing returns to scale. We may thus assume that the aggregate production function is a representation of an indeterminate number of firms that are in full competition.

The qualitative results of the model of course depend on the shape of F. Solow considers quite a number of different possibilities, but the one best remembered and usually quoted is when F has the Inada properties ($F_x \to \infty, 0$ as $x \to 0, \infty$ and $F\left(0,c\right) = F\left(c,0\right) = 0$). Because of the CRS assumption, we may write this model in *per capita* terms by dividing both sides of (2.12) by L and substituting (2.13) in. This leads to the differential equation

$$\dot{k} = sf\left(k\right) - nk$$

where lowercase variables are per capita, and $f\left(k\right) = F\left(K/L, 1\right)$. By the Inada assumption, f exhibits decreasing returns to scale, so that the equation has a single solution k^* to which all time-paths must converge. This implies that there exists a level K/L at which the extra capital only just compensates the increase in population. This is the steady state to which the economy converges, and in which the growth in production per capita

---

[16] The assumption of a fixed rate of saving can be relaxed without altering the basic results of the model. A model of intertemporal optimization was built by Cass (1965) and Koopmans (1965); the result may also be found in Barro and Sala-i-Martin (1995) and Rensman (1996).

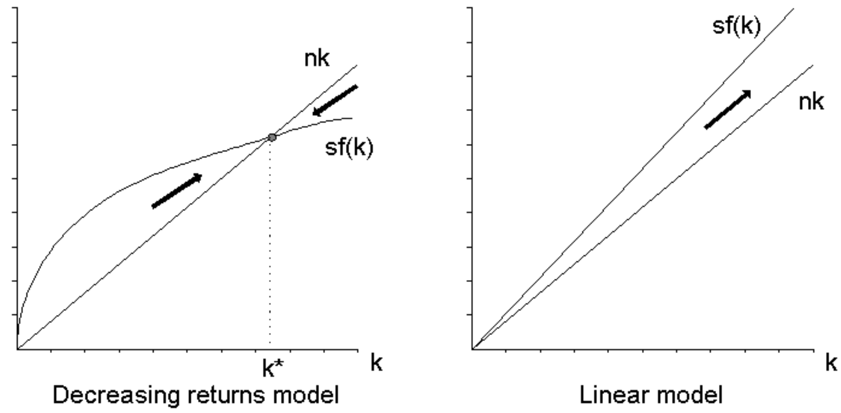

Figure 2.3: Direction of motion of k in two models of growth

stops. The model is depicted in the left-hand panel of Figure 2.3. Capital per worker converges to the steady state level k^* from every initial level k_0.

To stay in line with the empirical fact that the economy keeps growing, the neoclassical model is usually amended with exogenous technological growth. This growth is necessarily Harrod-neutral (for a proof, see Barro and Sala-i-Martin 1995, p. 54) and can be incorporated by substituting $\hat{L}_t$ for L_t in (2.12), with $\hat{L}_t = A_t L_t$. Regular increases in A then result in a growing income per capita, even if the economy is in the steady state. If the rate of growth of A is assumed constant it is possible to estimate values for it for different countries using time series data. In another paper, I estimated exogenous growth for the U.S. to be 0.0180 $[.0009]$ and for the Netherlands 0.0149 $[.0021]$ (standard errors in brackets, Knaap 1997).

The neoclassical model highlights the process of capital accumulation in a closed economy and does not consider the interactions between several economies. It does make a prediction about the dispersion of capital per head over several closed economies, if these economies can all be described by the same production and investment functions: regardless of the initial level of capital, the economies will converge to the same equilibrium, and thus to the same level of K/L. This property of the model is known as the convergence property.

The temporary nature of growth in this model has to do with the fact that the factors that can be accumulated together face decreasing returns to scale. The more of these accumulable factors are around, the less their added productivity is. This is an assumption of the model, and not necessarily a fact of life. The assumption was made because the neoclassical model also considers the factor labor, which cannot be accumulated by sheer economic means, and together the factors must exhibit CRS. For, if they do not exhibit CRS, the assumption of perfect competition is inappro-

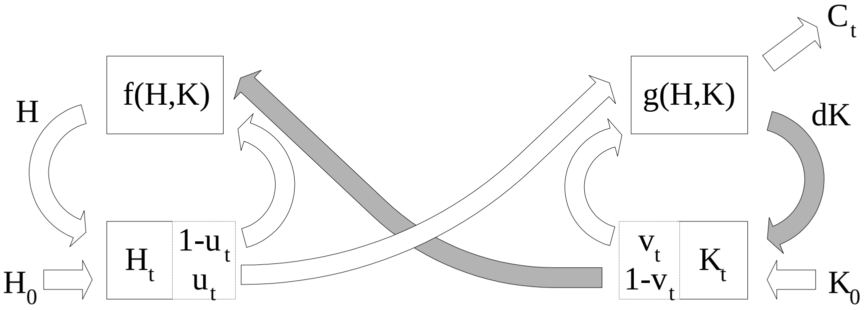

Figure 2.4: A box-arrow sketch of the two-sector model

priate.

On the premise that we will discuss the appropriate market structure below, let us now explore what would happen on a macro-level if *all* factors of production could be accumulated. This implies a return to the models proposed by Harrod (1939) and Domar (1946), who supposed that every addition to the stock of capital per worker allows production to be increased proportionally. Then the per capita stock of capital can never be too high, in the sense that additions to it are relatively unproductive. This can be seen when we substitute $F(K_t, L_t) = AK_t$ in formula (2.12) above. The accumulable resources in this case must be understood to include human capital and other production factors as well, besides capital in the narrow sense.

A graphical analysis of this linear model of production is in the right-hand panel of Figure 2.3. It is clear that if all factors can be accumulated, while the CRS condition still holds, we have specified a model of endogenous, ever-lasting growth.

An important point made by Rebelo (1991, p. 502) is that to achieve this result, not every part of the economy needs to have constant returns. It is sufficient that there exist a sector that uses a core of accumulable factors with a constant returns technology. This sector then becomes the economy's "engine of growth" as it pulls the rest of the economy.

We will illustrate this and other issues by considering the following two-sector model of an economy, taken from Barro and Sala-i-Martin (1995, p. 198):

$$C_t + \dot{K}_t + \delta_K K_t = A(v_t K_t)^{\alpha_1} (u_t H_t)^{\alpha_2} \tag{2.14}$$

$$\dot{H}_t + \delta_H H_t = B((1-v_t)K_t)^{\eta_1} ((1-u_t)H_t)^{\eta_2} \tag{2.15}$$

A box-arrow sketch of this model is in Figure 2.4. The different colors of the arrows are used later; for now consider them all equal.

We see that there are two sectors, one with production function f (formula 2.15) and one with production function g (formula 2.14). Both sectors use two factors, K and H. In principle, both factors K and H can be accumulated. The variables v, u and C are control variables. The sectors differ in parameters α_i, η_i, A, and B and in the fact that the sector that produces K also produces consumption, C. Consumers solve the dynamic problem

$$\max_{u_t,v_t} \int_0^\infty \frac{C_t^{1-\sigma}}{1-\sigma} e^{-\rho t} \mathrm{d}t$$

given H_0 and K_0 and the parameters.

The complete model (2.14)-(2.15) is analyzed by Mulligan and Sala-i-Martin (1993). They derive the conditions under which this system can generate a steady state growth path, that is, a solution path where all variables grow at a constant rate. It turns out that this is only possible under the following condition:

$$(1-\alpha_1)(1-\eta_2) = \alpha_2 \eta_1 \tag{2.16}$$

A model whose parameters do not obey this condition either comes to rest at equilibrium levels of H and K or 'explodes', which means that it generates infinitely large state variables in finite time, and the objective integral becomes improper. This knife-edge condition on the parameters bothered Solow (1994) who discusses the value of α_1 in the AK model (see below). If that parameter is only slightly different than assumed, condition (2.16) is not satisfied and the endogenous growth results vanish. It causes him to call this type of theory "unpromising on theoretical grounds" (p. 51).

The model (2.14)-(2.15) has a number of well known special cases. We briefly list them below.

Example 2.4.1 *The AK model. For this model, the sector on the left in Figure 2.4 is taken out. The other sector is assumed to have constant returns: $\alpha_2 = 0$, $\alpha_1 = v_t = 1$. Notice condition (2.16) is satisfied. This is a limiting case of the neoclassical Solow-Cass-Koopmans model with $f(K) = AK$, hence the name. The steady state solution is $\dot{C}/C = \dot{K}/K = (A - \delta_K - \rho)/\sigma$. The model does not have any transitional dynamics. The growth rate of C always remains positive under suitable parameters.*

Example 2.4.2 *The engine of growth. The two grey arrows in figure 2.4 are taken out. Both sectors have constant returns: $\eta_1 = 0$, $\eta_2 = 1$, $\alpha_1 = 1 - \alpha_2$, $\delta_K = \dot{K}_t = 0$. Here, K represents the invariant stock of non-reproducible, non-depreciating capital goods (think of land, for instance) and H is the stock of factors that can be accumulated. Again, the model only has a steady state solution and lacks transitional dynamics. Rebelo (1991) shows that the solution is $\dot{C}/C = \alpha_1 \dot{H}/H$ which is equal to $\alpha_1 (B - \delta_H - \rho) / (1 - \alpha_1 (1 - \sigma))$. It is natural to designate the sector producing H as the engine of growth, as it is the constant returns accumulation of H that causes C to grow.*

Example 2.4.3 *The Lucas model. This is a slightly more general version of the 'engine' model from Example 2.4.2, analyzed in Lucas (1988). This time we take out only the middle grey arrow. The parameters are* $\eta_1 = 0$, $\eta_2 = 1$, $\delta_{H,K} = 0$, $\alpha_1 + \alpha_2 > 1$. H *is understood to be human capital and* K *is conventional capital. Thus capital goods play no role in the (constant returns) creation of human capital. The goods sector shows increasing returns. In fact, Lucas assumes constant returns plus an external effect of the average stock of human capital, so that a competitive equilibrium exists (more on this below). The optimal steady state growth rate of consumption (with zero population growth) is* $\dot{C}/C = \dot{K}/K = (\frac{1-\alpha_1+\gamma}{1-\alpha_1}B - \rho)/\sigma$. *Here,* $\gamma = \alpha_1 + \alpha_2 - 1$, *the size of the external effect. This shows that increasing returns are not essential for the resulting endogenous growth, as* $\gamma = 0$ *still permits a positive value for* γ_C.

These models can be classified as to their stability. Because there are no transitional dynamics in the first two models, a small perturbation of the initial value has lasting effects. Because the growth rate of the accumulable factor is constant, the difference between the solutions starting in F and $F + \varepsilon$ grows exponentially (F is the initial value of the relevant state variable, K and H for the AK and the 'engine' model, respectively). A similar result holds for the Lucas model, although derivation of this result is not trivial. See Barro and Sala-i-Martin (1995, p. 184) and Mulligan and Sala-i-Martin (1993, p. 758).

The micro level

The models presented above pose a difficulty additional to the knife-edge condition on the parameters. If they include increasing returns to an accumulable factor, the usual fully competitive environment is no longer feasible; in other words, the set of supporting prices does not exist. We look at two approaches that have been used to circumvent this problem. One is to introduce increasing returns only at the level of the sector, and not of the firm. The sectorial returns take the shape of externalities. The other approach is to explicitly model the imperfect competition that arises because of the increasing returns.

Externalities We discussed externalities in Section 2.2.1 as a means to reconcile CRS and increasing returns. Some endogenous growth models use non-pecuniary externalities to do just this. We have already mentioned the use of externalities in the Lucas (1988) model, and we now look at the approach in Romer (1986). Because of a careful specification of the externality setup, the model does not suffer from the knife-edge condition (2.16).

The production function for a representative firm is $F(k_i, K, \mathbf{x}_i)$ with k_i the state of knowledge available to firm i and $\mathbf{x}_i$ a vector of additional factors (capital, labor). The variable K is the aggregate level of knowledge

$\sum_{i=1}^{N} k_i$ which can be used by all firms to some extent because knowledge is partly non-rival and non-excludable. It is assumed that F has constant returns to the factors k_i and $\mathbf{x}_i$, and increasing returns to all three factors. However, each firm takes the value of K as given when making its decisions. Output can be consumed or invested in k_i ($\mathbf{x}_i$ is constant). The latter goes through the knowledge production function: $\dot{k}/k = g\,(I/k)$. The function g is increasing and bounded from above by a finite constant M. These conditions on g prevent the 'explosion' that the models above suffered from: a firm can never let its stock of knowledge grow at a faster rate than M so that k_i and K cannot reach infinity in finite time. Note that the g-functions above were usually linear in the state variable.

Romer finds that the socially optimal solution is different from the competitive solution because the latter does not take the external effects into account. Both solutions do generate endogenous growth, albeit that the rate of growth is larger in the optimal solution. The competitive solution is properly defined in all models that satisfy the above specification.

Monopolistic Competition As an alternative to the use of externalities above, Romer (1987) explicitly introduces markets that are monopolistically competitive; the model is very similar to that in Section 2.2.3. There exists an all-purpose capital good Z, which is transformed into a continuum of n^* intermediate goods; this is done by a continuum of firms (see appendix 2.A). These intermediate goods are then used as inputs for the final good. The final good can again be added to Z or can be consumed. Consumers maximize utility (a function of consumption) intertemporally. The production function in the final goods sector is as described in section 2.2.2. An increasing number of intermediate inputs (n^*) increases output as in the example in Section 2.2.3. Varieties $x\,(i)$ are produced using an increasing returns production function.

The most important characteristic of this model is that output Y turns out to be a linear function of the stock Z. This is because the efficient scale of the intermediate producers does not change as Z changes, so n^* is linear in Z. As Y is linear in n^* this means that the model behaves much as though it were the AK model above, and generates stable endogenous growth. It also suffers from the above-mentioned drawbacks, notably the fact that it is parameter-unstable. However, constant returns of Y to Z seem a little less "luck" (cf. Solow 1994, p. 51) than above, as they can be defended on economic grounds rather than just being mathematically convenient. Also, this model became the backbone of more advanced growth models. We will come across those models in the next section.

2.4.2 Growth through innovation

Above, economic growth was mostly brought about by an ever increasing supply of factors. In Romer's (1987) model, an increase in the number of varieties played a role, but this increase was 'free,' *i.e.* no sacrifices needed to be made to discover the new varieties; the increase was a matter of efficient scale. Yet stylized fact #4 (footnote 15) specified that 'technological advance comes from things that people do.' The second wave of growth models thus concentrated on a situation where R&D absorbs resources and new varieties are discovered in return.

New varieties can be substitutes or complements to older ones. In growth theory parlance one thus distinguishes *horizontal* and *vertical* innovation. The term 'horizontal innovation' is from Grossman and Helpman (1991), and the first model in this direction was drafted by Judd (1985). It is replicated here.

It is assumed that consumers maximize an intertemporal CES utility function

$$U = \int_0^\infty e^{-\beta t} \left(\int_0^{V(t)} x\left(v,t\right)^\theta \,\mathrm{d}v \right) \mathrm{d}t. \tag{2.17}$$

The only factors of production are labor, which is constant at L, and the known range of varieties $V\left(t\right)$. For each variety there holds that one unit can be produced using one unit of labor. The range of varieties $V(t)$ grows through R&D, whose only input also is labor. It is assumed that $\dot{V} = L_{\text{R\&D}}/k$.[17]

There holds that $\theta < 1$, so that in equilibrium the quantities $x\left(v\right)$ are the same across varieties. Call this quantity y. The problem may then be written as

$$\max_{0 \le y \le LV^{-1}} \int_0^\infty e^{-\beta t} y^\theta V \mathrm{d}t$$
$$\text{subject to } k\dot{V} = L - yV.$$

The solution (see Judd 1985) is that the economy converges to a stationary state where both y and V are constant. That is, there exists an optimal variety of goods, and once this variety has been attained innovation comes to a halt.

It is possible to see why innovation stops if we compare the problem to the basic monopolistic competition model of section 2.2.2. In that setup, an increase in that number of firms lowers each firm's profit margin. Profit is used to repay a fixed cost that is associated with entry. A situation of too many producers leads to profits that are too low to recoup the initial investment. Hence there exists an optimum number of producers. In this

---

[17] Note that there is no uncertainty involved in research. This rather quaint assumption is maintained through much of the growth-through-innovation literature.

model, the fixed cost associated with entry is the labor that must be hired to conduct R&D. If that cost cannot be repaid because profit margins are too low, innovation stops.

Note that the MC market form is essential in this model because, as opposed it full competition, it allows producers to make a profit. Those profits can be used to pay off the initial R&D expenses. Without the possibility to price higher than marginal costs, innovation would never occur.

Horizontal innovation, endogenous growth

One way to keep the economy growing in the model above, is by lowering the costs of innovation as the number of varieties increases. If the outcome of the model should be a constant growth rate g of the number of varieties, and we know that $\dot{V} = L_{\text{R\&D}}/k$, then we can deduce

$$\begin{aligned} \frac{\dot{V}}{V} &= g = \frac{L_{\text{R\&D}}}{kV} \\ g \text{ constant} &\Rightarrow kV \text{ constant} \end{aligned}$$

So, if the R&D productivity parameter k^{-1} is proportional to the number of varieties V, we can have everlasting growth.

Romer (1990) presents an adapted version of his model in Romer (1991) that "emphasizes the importance of human capital in the research process" (p. S78). Like above, it features three sectors: R&D, intermediate and final goods. Knowledge has a rival component H and a non-rival component A; the latter can be interpreted as the 'state of technology' and is allowed to grow without bounds.

In the production of final output Y, human capital H plays a role next to labor L and a continuum of intermediate goods $x(i)$:

$$Y(H_Y, L, x) = H_Y^{\alpha} \cdot L^{\beta} \cdot \int_0^A x(i)^{1-\alpha-\beta} \mathrm{d}i \tag{2.18}$$

(notice the similarity to formula 2.17 above). The stock of H is split up in a part H_Y that works in the final goods sector, and a part H_A that works in the R&D sector. The interval over which $x(i)$ is positive has size A, the level of technology. An increase in A, that is, a rise in the level of technology, does not render older types of the intermediate good obsolete. This is due to the additively separable form of (2.18).

In line with the derivation above, the technology used in the R&D sector is such that A changes according to

$$\dot{A} = \delta \cdot H_A \cdot A \tag{2.19}$$

This form is justified by claiming that a larger stock of knowledge will enhance current research possibilities. The model is closed by specifying that the stock of intermediate goods $K = \int_0^A x(i)\mathrm{d}i$ evolves according to $\dot{K}_t = Y_t - C_t$.

Romer's analysis shows that the model, specified above, yields unbounded endogenous growth. This is caused by the assumption of constant returns to scale in equation (2.19) above. With respect to this assumption, Romer writes:

> "...in this sense, unbounded growth is more like an assumption than a result of the model. [...] Whether opportunities in research are actually petering out, or will eventually do so, is an empirical question that this kind of theory cannot resolve."

Vertical innovation, endogenous growth

Aghion and Howitt (1992) consider a model of growth that features vertical innovation. Newer types of intermediates replace the older types, and therefore the model represents the concept of Creative Destruction introduced by Schumpeter (1942).

The economy consists of three sectors: the R&D sector, the intermediate goods sector and the sector that produces consumption goods. The tradeoff in the economy is the decision how many workers are allotted to work in R&D instead of the intermediate goods sector. This number depends on the expected profitability of innovations.

A new intermediate good completely replaces the older type. The inventor is the only producer of the goods, and is thus allowed to earn some monopoly rents until the next innovation takes place. The time until the next innovation is random and exponentially distributed, and depends negatively on the number of people working in R&D. The marginal product of an extra R&D worker is decreasing, so that there exists an optimal number of people engaged in research and development.

There is only one kind of uncertainty in the model, namely the time of arrival of a new technology. The increase in the level of technology, caused by the invention, is fixed. By defining a 'period' as the elapsed time between two innovations, the authors in effect make the monopoly rent earned off the inventions the random variable in the model.

Without being explicit about such things as the aggregate production function, Aghion and Howitt (1992) examine the motives for investing in R&D and find that, depending on the 'arrival function' of new technologies, there may exist a fluctuating or steady (possibly zero) number of researchers in the economy. Endogenous growth is implied as soon as there is a positive number of researchers active, and its rate is determined by both endogenous and exogenous variables.

2.4.3 Empirical tests

In this section we will mostly look at empirical tests of the implications of the above models. As some of the results came out negatively, interest in the neoclassical Solow model was revived in the early 1990s. The results of such interest can be found in Mankiw et al. (1992) and Nonneman and Vanhoudt (1996).

As Pack (1994) notices, much early empirical research on endogenous growth models is conducted in the neoclassical framework. Thus, instead of testing the new growth theory directly, it is only used as a possible alternative when the Solow model fails. A first direct test of the theory is performed by Jones (1995b), who tests the time-series predictions of new growth theory. The evidence is collected in two rounds.

The models of this section have the property that a permanent increase in investment causes a permanent increase in the economy's rate of growth. Or, even stronger, the two variables are linearly related. This is easily seen with the AK and the Lucas model, as the rate of growth of capital and the rate of growth of consumption are the same. The result does not hold for the engine-of-growth model. Jones (1995b, p. 500) shows that the growth rates of selected OECD countries are stationary variables, whereas a unit root in the OECD investment rates can only be rejected in four out of 15 cases. Almost all investment rates show a positive trend. This contradicts the (supposed) linear relationship between investment rates and growth rates. Further time series estimations show that the effects of an increase in the investment rate can only be observed for eight years after the shock, much less than the proposed everlasting effect.

The testable proposition of the R&D-based models of section 2.4.2 is that the growth rate of an economy is linearly related to the number of people active in the R&D sector. Using data on the number of researchers in the U.S., Germany, Japan and France, Jones (1995b, p. 517) again shows a strong upward trend in these explanatory variables, whereas the rate of growth of their respective countries remains stationary. These two results can be seen as a rejection of the testable propositions that came out of the endogenous growth models. Jones (1995a) proposes to 'fix' the R&D-based model by writing equation (2.19) in Romer's (1990) model as

$$\dot{A} = \delta \cdot H_A^{\lambda} \cdot A^{\phi}$$

If $\lambda, \phi < 1$ then the model will no longer exhibit endogenous growth but instead settle down in an equilibrium. As Jones (1995a, p. 766) puts it, "…$\phi = 1$ represents a completely arbitrary degree of increasing returns and […] is inconsistent with a broad range of time series data on R&D and TFP growth" (see also Romer's quote in section 2.4.2). The model proposed by Jones (1995a) can best be seen as an extended version of the Solow (1956) setup, with all its asymptotic characteristics.

2.4.4 Review

We have seen that classical growth models that use the CRS paradigm explain growth through accumulation, but this growth cannot last forever without exogenous propelling. Accumulation-based models can explain lasting growth if they have constant returns to all accumulable factors. The micro-foundations for these models use externalities or an MC-setup.

The second wave of endogenous growth models explains growth not by accumulation of factors, but by technological progress. Virtually all these models use the MC framework.

Some critical notes can be placed about endogenous growth models. The scale-effects that they predict are not observed, and they are parameter-unstable. Despite the critical notes above, at the time of writing, endogenous growth theory is still very much alive. It turns out that the spirit of the models can be maintained while accommodating empirical facts (see Aghion and Howitt 1998, chapter 12). And the ability of the models to handle a number of questions that exogenous growth theory cannot answer (questions concerning the long run growth rate, for instance) has made them popular with empirical researchers.

2.5 Dynamic economic geography

This review chapter has shown how the monopolistic competition model (section 2.2) made possible new ways of modelling economic geography (section 2.3) and economic growth (section 2.4). So far however, we have only discussed growth models without an explicit geographical dimension, and mostly static geography models.[18] It is only natural to combine the two strands of the literature, which are based on the same framework.

Recently, a large amount of research was done in this direction. A good survey of these efforts is in Baldwin and Martin (2003). They discern two classes of models, one of which we have already discussed above: models in which growth affects geography, but not the other way around. The other class of models allows for an interplay between growth and agglomeration; we survey it in this section.

The key assumption that determines whether growth and geography will interact concerns the (spatial) range of knowledge spillovers. As we have seen above, spillovers are crucial to endogenous growth theory. They imply that the stock of productive knowledge that is already available in the economy helps to reduce the costs of acquiring extra knowledge. An assumption about spillovers is implicit in capital-accumulation equations

---

[18] The exception being the CP-model in which capital accumulation is the medium for complementarities on page 29.

such as (2.19) above. There, the size of the available non-rival knowledge stock A determines the effort that is needed to increase it.

It is intuitive that whatever we assume about the geographical reach of this effect will determine the growth performances of different regions. If local knowledge only spills over to R&D efforts in the same region, we can imagine one region growing, while the other stagnates. When spillovers are worldwide, as we assumed above in formula (2.11), different regions can pool their knowledge.

In the next paragraph, we discuss the results when spillovers are local. We further survey a number of other approaches with different agglomeration links. Paragraph 2.5.2 discusses a model by Martin and Ottaviano (1996b) where the linkage runs through the R&D sector, in a way reminiscent of Krugman and Venables (1995). Paragraph 2.5.3 shows that the linkage can go through the (research-) labor market as well. We end with some other models in paragraph 2.5.4.

2.5.1 Agglomeration through local knowledge spillovers

Baldwin and Martin (2003) replace the capital accumulation equation in formula (2.11) with a slightly different version that involves transport costs for knowledge.

$$\dot{K}_i = \gamma \cdot L_i^I \cdot (K_i + \lambda \cdot (K_{\text{world}} - K_i)) \tag{2.20}$$

The parameter that measures knowledge transport costs, λ, lies between zero (only local spillovers) and one (global spillovers). Its value determines the rate of accumulation of region i's capital, $\dot{K}_i$. For now, we assume that $\lambda = 0$.

As before, we discern two cases with respect to capital mobility. Perfect mobility means that agents can own capital in every region; this equalizes the rate of return to capital in all regions. Capital immobility means that agents can only accumulate capital in their own region. This assumption applies to human capital, for instance, if people are not mobile.

Perfect capital mobility

When there are no restrictions on owning foreign capital, everybody will want to produce capital where the costs of production are lowest. According to (2.20), the region with the highest initial capital stock is where all capital production will take place. The lagging region does not have an innovation sector of its own, but is perfectly able to accumulate capital using the other region's knowledge pool. This means that there is no self-enforcing agglomeration in this model, as an initial disadvantage does not translate into lower income for the lagging region.

There exists an interesting trade-off in this model. Assume for a moment that region N has a larger share of total income, due to a higher initial share of world capital. The share of firms that locate in N will be higher than N's share of income, due to the home market effect (see appendix 2.B). This concentration of firms leads to a higher rate of growth, because of the larger local spillovers that now occur in region N.

How does this impact the people in the S-region? On the one hand, a larger share of income going to their neighbors in N is bad, and more firms locating in N means higher transport costs for the people in S, who will have to import more. But on the other hand a higher rate of growth benefits inhabitants of all regions. The balance of these two effects is determined by the costs of transport for knowledge and goods. Baldwin and Martin show that for small λ (when spillovers are mostly local) and small costs of transport for goods, the S region may *gain* from an extra concentration of industry in N. This is a positive and surprising outcome, that contrasts with the earlier result that agglomeration in one region is generally welfare-reducing for the other region.

Immobile capital

When capital is immobile, agents can only invest in their own economy. If the returns to capital in one region drop, this has an immediate effect on the income of the (capital owning) inhabitants of that region, leading to a feedback that may cause the forming of a periphery and an agglomeration. Earlier, on page 29, we stated that a model with global spillovers and immobile capital could generate a core-periphery outcome. It should not be surprising that the same holds for a model with immobile capital and local spillovers.

The extent to which an inequality between the rate of growth in two regions arises, depends on the level of transport costs. Baldwin and Martin (2003) show that there exist a threshold level of transport costs below which a process of agglomeration starts. When transport costs are high enough, local demand makes investment worthwhile in either region and a symmetric equilibrium obtains. However, when transport costs drop below the threshold level, one region completely stops investing while the other experiences a 'growth takeoff' (p. 28). The region that agglomerates sees the costs of investment fall more quickly due to local spillovers, and enters a period of high growth. The other region gets stuck in a situation where local demand is too low to justify investment in new firms.

> Hence, growth affects geography which itself affects growth and agglomeration is driven by the appearance of growth poles and sinks. (Baldwin and Martin 2003, p.28)

The question once again arises whether the region where innovation stops is worse off; the higher rate of growth is beneficial to both regions, after all. As it turns out, the welfare of the region that was left behind depends on the share of differentiated goods in expenditures. For a high value of this share, the region may actually gain from the new, agglomerated, equilibrium. For low values, it certainly loses.

2.5.2 Agglomeration through the R&D sector

Localized spillovers are not the only way in which growth and geography may interact. The model in this section has a feedback between growth and agglomeration that is the result from vertical linkages between the R&D sector and the differentiated goods sector.

Martin and Ottaviano (1996b) present a model with two regions and three sectors: a full-competition agricultural sector, an MC industrial sector and a sector for R&D. It is the latter sector that is most interesting.

The R&D sector is fully competitive. The output of the sector is patents; each patent can be used to manufacture a variety in the industrial sector, the total number of varieties is n. The productivity of the R&D sector increases as n gets larger. These qualities are similar to the Romer-Grossman-Helpman models of Section 2.4.2. The only input to the R&D sector is the composite good D that is the output of the industrial sector. This creates a linkage between the two sectors akin to the linkages in Krugman and Venables (1995) and Venables (1996a). Wherever firms from the industrial sector are abundant, the costs of R&D are low. And wherever R&D is conducted, the demand for industrial goods is higher. The linkage causes agglomeration of industrial and R&D firms in the same location.

Consumers in this model maximize an intertemporal utility function that depends on the consumption of the agricultural and the industrial good. The model has two types of solutions. In one solution, both locations have exactly the same number of industrial producers. R&D is conducted in both locations. This solution is unstable. The other solution has all R&D taking place in one location, where also the majority of the industrial producers are active.

In the second, unbalanced, solution the rate of growth is higher. This is intuitive: if industrial producers are spread evenly the industrial composite costs are the same in both locations, say, c. In case of an imbalance, there always is a location in which the composite is cheaper than c. Because R&D uses only the composite, an even spread of the industrial producers maximizes production costs and minimizes growth.

An important conclusion of the model by Martin and Ottaviano (1996b) is that the rate of growth influences the location decision, and the location decision influences the rate of growth. This puts models in which both are treated separately at a disadvantage. The fact that the interaction causes agglomeration of industrial activity is in line with the quote by Myrdal on page 12.

2.5.3 Agglomeration through the labor market

Migration between the two regions is the cornerstone of the model by Baldwin and Forslid (1997), just as it is in Krugman (1991a, 1991b).

The assumptions are roughly the same as above, except that the R&D sector now uses only labor as an input. Again, the input requirement decreases as the stock of knowledge gets larger. However, the spillovers are only regional; the stock of knowledge consists of the number of firms in ones own location only.

In the long run, the linkage now works as follows: wherever the most firms are is where the consumer price index is lowest. Personnel has an incentive to move to this location. So do all firms in the (competitive) R&D sector, because the costs of R&D depend negatively on the available pool of knowledge (in this case, the number of firms). On the other hand, where most people are is where firms like to be because of the demand that people exercise, and because of the larger labor market that the firms can draw from.

Again, there are two types of equilibrium in this model. In one, all activity is evenly divided between locations, and both locations grow at the same speed. The other equilibrium has all R&D and most labor and industrial firms in one location, the other deprived of most activity.

It turns out that the first equilibrium (the even spread) is very unstable, even at prohibitive trade costs. This was not the case in the Krugman (1991a, 1991b) models. Contrary to the static economy, the dynamic economy will agglomerate into one location for all possible parameters.

In this model, the R&D sector does not constitute a part of the linkages, as it did above. However, it does react to the outcome. In the long term, all R&D is concentrated in the agglomeration, because it is the cheaper place to work. This does not necessarily affect the location of the industrial firms developed by the R&D sector, as the patents are valid in both locations. Thus, the R&D sector reacts to the linkages, but is not a part of it.

The interplay between growth and location shows up in this model as well. When all R&D is done in the same location, all R&D firms add to the same stock of knowledge. This leads to faster rates of growth than if the advances are divided over two separate stocks of knowledge, because the efficiency of the R&D sector increases with K.

2.5.4 Other models of growth and geography

We discuss a number of other models where growth and location theories are integrated. The degree of interaction between the two is more limited than in the theories that were discussed above.

Quah (2002) proposes a highly theoretical model where the extent to which spillovers between different regions exist is a function of the distance between those regions. The advantage of his approach is that space is no longer limited to two regions, but can consist of a continuous plane, or a globe. The model shows that if there are adjustment costs for capital and spillovers are local, there exists an interesting transition path to the long run equilibrium, where all regions are equal. During the transition, an agglomeration force creates growth 'peaks' and 'troughs' in the space that is studied. These poles disappear when the long run steady state is attained.

Martin and Ottaviano (1996a) develop a model where migration does not occur. There are three sectors, agriculture, industrial and R&D. The MC industrial sector uses patents as in Section 2.4.2. The competitive R&D sector that develops the patents uses labor and the pool of knowledge. Patents can be used in any location and are not subject to transport costs. If the R&D sector has access to all knowledge in the economy (global spillovers), then R&D is conducted in both locations. If there are only local spillovers, the R&D sector agglomerates. The developed firms will be set up in both locations, though.

The model is a first attempt to merge theories of growth and location. The structure of the economy (industrial and agricultural production at the two locations) is so rigid that it does not change much under the different growth regimes, so that the interaction is limited to the location of the R&D sector.

Englmann and Walz (1995) construct a model with two locations without transport costs. The geographic structure plays a role however, because the knowledge pool is different between the two regions. This leads to a situation with nontraded inputs, where each location has its own intermediates. The initially larger region becomes the industrial center, whereas the other becomes a peripheral region. If there are interregional knowledge spillovers, so that inputs still are not traded but R&D *can* use them, many solutions become possible.

In this model, devoid of transport costs, it is the size of the knowledge pools that steers the regional development. Knowledge pools contain nontraded inputs, so that the factor that causes agglomeration is not traded itself. Though interesting, this is fundamentally different from the models of Section 2.3 and is the subject of another branch of literature (see, for instance Rivera-Batiz and Romer 1991) .

Duranton and Puga (2003) include a section on dynamic externalities that lead to growth as well as agglomeration. The payoff to investment in human capital is thought to be a positive function of the human capital stock in the immediate vicinity. External effects of other people's human capital fuel growth, as they are a particular variety of the endogenous growth models from section 2.4. At the same time, they are an agglomerating force.

Redding and Schott (2003) connect geography and growth indirectly as they look at the effect of remoteness on the accumulation of human capital. As we have seen in section 2.4, the accumulation of human capital is thought to be a mechanism for economic growth. The authors find that, under plausible assumptions, remoteness depresses the skill premium and reduces incentives to accumulate human capital. Though their model is static and yields no direct results pertaining to the rate of growth, this indirect evidence points to a negative relation between the latter and the geographic position of a country.

2.6 Conclusions

In this survey paper, we introduced the monopolistic competition framework as the foundation of two new strands of literature, on the one hand endogenous growth theory, and on the other hand economic geography. Both theories use the fact that MC allows scale economies to be used in a model of general equilibrium.

In our survey of endogenous growth, we showed that early models were based on the endless accumulation of resources, as are exogenous growth models. Later versions stressed technological progress as the source of growth. Progress can take the form of horizontal innovations and vertical innovations.

In the literature on economic geography, linkages between firms and consumers, and between firms themselves, play an important role. The different models can be classified as to the type of linkage they use. Most models predict a dramatic agglomeration at certain parameter values.

Because both strands of literature rest on the same foundation, and describe related phenomena, it is only logical to incorporate the two. We surveyed several attempts to that end. It turns out that the interplay between growth and location upsets the predictions of either literature by itself. Stable equilibria in static geography models turn out to be unstable in a dynamic context; the rate of growth again is influenced by the location pattern, which depends on initial values.

Studies that investigate the empirical value of both literatures are not overly enthusiastic. Whereas CRS-based theory stands up to the data in a reasonable way, many effects predicted by MC are not measured at all. However, this may be due to a lack of testing methodology capable of dealing with the nonlinear nature of the models. Tests can only be conducted on specific linear predictions of the model. It is unclear to which extent a refutation of such a prediction constitutes a problem for the whole body of theory.

It seems that the combination of endogenous growth theory and economic geography is a promising field of research. The scattered results available so far indicate that more work needs to be done before any swaying conclusions can be drawn.

2.A A continuum of goods

The derivation of the equilibrium in the monopolistic competition framework holds in general 'when n is large.' This can be an awkward assumption; do we really need, in economic terms, an endless array of goods to work with this model?

The usual interpretation is that really, all we need is to be able to refine and differentiate goods enough. The range can remain the same, but we ought to be able to divide goods into as many different subtypes as we need. Mathematically, this means that we look at a continuum of goods $x\left(j\right)$ defined on a real interval $[0,n]$. In principle, each good $x\left(j\right)$ with $j \in [0,n]$ can be identified as a different variety. Quantities of goods, however, are only defined over intervals of j. The quantity $x\left(3\right) = 1$ is meaningless, but $x(j) = 1$ for all $j \in [0.1, 0.2]$ is a positive quantity.

How do our maximand U and the budget restriction change when we work with a continuum of goods? They can be derived as limiting cases of their discrete versions.

Suppose we call all the goods $x(j)$ with $0 \leq j < n_1$ good 1, all the goods with $n_1 \leq j < n_2$ good 2, and introduce a set of numbers $\mathcal{S} = \{n_0, n_1, n_2, \ldots, n_Q\}$ like this, with $n_0 = 0$ and $n_Q = n$. If two goods belong to the same interval, they are purchased in the same amount and priced the same.[19] With this set, we are back in the discrete goods setup. There holds

$$\begin{aligned} U &= \left[\sum_{i=1}^{Q} \left(n_i - n_{i-1}\right) x(i)^{\theta}\right]^{\frac{1}{\theta}} \\ E &\geq \sum_{i=1}^{Q} \left(n_i - n_{i-1}\right) x_i p_i. \end{aligned}$$

For any properly defined set $\mathcal{S}$, these formulae can be rewritten as

$$U = \left[\int_0^n x(i)^{\theta} \mathrm{d}i\right]^{\frac{1}{\theta}} \quad (2.21)$$

$$E \geq \int_0^n x\left(i\right) p\left(i\right) \mathrm{d}i \quad (2.22)$$

We see that there are two ways in which the number of goods can increase. By picking a larger set $\mathcal{S}$, we refine the definition of the goods, and allow for

---

[19] That is, we have $x(i)$ and $x(j)$ with $n_{k-1} \leq i < n_k$ and $n_{k-1} \leq j < n_k$, and both are purchased in the amount x_k.

more price and quantity differentiation. By increasing n, the range of goods is increased with the introduction of new varieties that can be purchased *instead of* the older set.

The monopolistic competition setup is usually introduced as in formulas (2.21) and (2.22), without a specific set $\mathcal{S}$ defined. To retrieve the results that hold in the integer case, however, we need to imagine such a set ourselves.

Suppose we want to maximize function U from formula (2.21) under the restriction (2.22). The problem can be written as a Lagrangian,[20]

$$\max_{\{x(i)|i\in[0,n]\}} \left[\int_0^n x(i)^\theta \mathrm{d}i\right] - \lambda\left[\int_0^n x(i)\,p(i)\,\mathrm{d}i - E\right]$$

The problem is hard to solve when we stick with the integral notation, but we can imagine that the differentiation between goods only goes as far as a set $\mathcal{S}$, which we do not specify. We may then write the maximand as

$$\mathcal{L} = \left[\sum_{i=1}^{Q}(n_i - n_{i-1})\,x(i)^\theta\right] - \lambda\left[\sum_{i=1}^{Q}(n_i - n_{i-1})\,x(i)\,p(i) - E\right].$$

Differentiate with respect to $x(i)$ and set equal to zero to find

$$\begin{aligned} \beta\,(n_i - n_{i-1})\,x(i)^{\theta-1} - \lambda\,(n_i - n_{i-1})\,p(i) &= 0 \Rightarrow \\ \theta x(i)^{\theta-1} &= \lambda p(i)\,. \end{aligned}$$

Note that we may divide by $(n_i - n_{i-1})$ because the requirements for $\mathcal{S}$ have it greater than zero. Because λk does not vary with i, we may write that for all i,

$$x(i)p(i)^{\frac{1}{\theta-1}} = \text{constant.}$$

If we substitute this into formula (2.22) we get that

$$x(i) = \frac{Ep(i)^{-\sigma}}{\int_0^n p(j)^{1-\sigma}\mathrm{d}j}$$

where $\sigma = 1/(1-\theta) > 1$.

2.B The home market effect

We follow the derivation in Neary (2001) in this section. Assume that the two regions are exactly equal and consider an equal and opposite change in the environment. That is,

$$\begin{array}{llll} p = p^*, & q = q^*, & n = n^*, & E = E^* \\ \hat{p} = -\hat{p}^*, & \hat{q} = -\hat{q}^*, & \hat{n} = -\hat{n}^*, & \hat{E} = -\hat{E}^* \end{array} \tag{2.23}$$

---

[20] We momentarily omit the exponent $1/\theta$, which does not change the outcome of the maximization.

where the *foreign* region has asterisks on its variables and a 'hatted' variable denotes a rate of change, *i.e.* $\hat{x} = \mathrm{d}x/x$. We will use equations (2.8) and (2.9) to derive the result, starting with the latter which we replicate here for convenience.

$$q^{1-\sigma} = np^{1-\sigma} + n^* \left(\frac{p^*}{\tau}\right)^{1-\sigma} \tag{2.9}$$

We totally differentiate equation (2.9), which gives us

$$(1-\sigma)q^{-\sigma}\mathrm{d}q = p^{1-\sigma}\mathrm{d}n + (1-\sigma)p^{-\sigma}\mathrm{d}p + \left(\frac{p^*}{\tau}\right)^{1-\sigma}\mathrm{d}n^* + n^*(1-\sigma)\left(\frac{p^*}{\tau}\right)^{-\sigma}\frac{\mathrm{d}p^*}{\tau}$$

Using the 'hat'-notation and the equalities from (2.23), we can write this as

$$(1-\sigma)q^{1-\sigma}\hat{q} = (\hat{n} + (1-\sigma)\hat{p})(1-\tau^{\sigma-1})np^{1-\sigma}$$

while we can rewrite (2.9) as

$$q^{1-\sigma} = np^{1-\sigma}(1+\tau^{\sigma-1})$$

which combines into the first result,

$$\hat{q} = Z\left(\frac{1}{1-\sigma}\hat{n} + \hat{p}\right) \tag{R1}$$

where $Z = \frac{1-\tau^{\sigma-1}}{1+\tau^{\sigma-1}}$.

Next, we use equation (2.8) to write the total demand for a firm, which is simply the sum of home and foreign demand:

$$x_i = p_i^{-\sigma}\left(Eq^{\sigma-1} + E^*(q^*)^{\sigma-1}\tau^{\sigma-1}\right) \tag{2.24}$$

which, after total differentiation and use of hats, gives

$$\begin{aligned}\hat{x}_i &= -\sigma\hat{p}_i + \frac{p_i^{-\sigma}}{x_i}\left(\hat{E} + (\sigma-1)\hat{q}\right)Eq^{\sigma-1} + \\ &\quad \frac{p_i^{-\sigma}}{x_i}\tau^{\sigma-1}\left(\hat{E^*} + (\sigma-1)\hat{q^*}\right)E^*(q^*)^{\sigma-1} \\ &= -\sigma\hat{p}_i + (1-\tau^{\sigma-1})(\hat{E} + (\sigma-1)\hat{q})\frac{p_i^{-\sigma}Eq^{\sigma-1}}{x_i}\end{aligned}$$

where we invoked the equalities from (2.23) in the second step. We now rewrite formula (2.24) as

$$x_i = p_i^{-\sigma}(1+\tau^{\sigma-1})Eq^{\sigma-1}$$

which, combined with the above, gives the second result

$$\hat{x}_i = -\sigma\hat{p}_i + Z(\hat{E} + (\sigma-1)\hat{q}). \tag{R2}$$

Now suppose that from the symmetric equilibrium, for some reason, the home region faces an increase in demand. The MC model tells us that per-firm output is fixed at its optimal level, as is the price of both types of goods. This means that an increase in demand leads to the entry of new firms into the home market. Because these new firms produce new varieties, this changes the price index of industrial goods which may lead to additional entry or exit of firms. The results derived above allow us to quantify the effect of a demand shock on the number of firms.

In the above notation, the increased demand means that $\hat{E} > 0$. We know prices and per-firm output do not change, which gives $\hat{p}_i = \hat{x}_i = 0$. From (R2) we find that $\hat{E} = (1 - \sigma)\hat{q}$. Using this to substitute $\hat{q}$ from (R1), we find that

$$\hat{E} = \frac{1}{Z}\hat{n}.$$

From the definition of Z above, we know that $0 < Z < 1$ if transport costs are positive. This allows us to interpret the above formula as the *home market effect* (Krugman 1980): the region with a higher demand has a proportionately higher share of manufacturing.

2.C The break and sustain point

To find the break point, the level of transport costs at which the symmetric equilibrium becomes unstable, we start at the same point as the previous paragraph. Suppose there are two regions and the symmetric equilibrium has been attained. That is, equation (2.23) is in force. We again use the notation $\hat{x} = \mathrm{d}x/x$.

We assume that workers are mobile between regions and move to equalize the level of real wage $\omega = wq^{-\mu}$. Firms enter and exit to drive profits to zero, but in this derivation we assume that this proces of adjustment is much slower than that of the (mobile) workers. This assumption allows us to assume that the equalization of real wages holds at all times. We then study the properties of the equilibrium by looking at the direction of change in firm profits, in response to a change in the number of firms. If we had reversed the assumption about the speed of adjustment, we would hold profits at zero and look at the change in real wages. Puga (1999) shows that the results of either assumption are the same.

We gather some relationships between rates of change in the model. From (2.5), the pricing rule, we find that $\hat{p} = \hat{w}$. Condition (2.6) implies that $\hat{L} = \hat{n}$. Finally, the real-wage condition (2.10) renders us $\hat{\omega} = \hat{w} - \mu\hat{q}$.

Total expenditure in a region is the sum of the total wages of the industrial and agricultural workers, or $E = wL + L_A$, where wages for the agricultural sector are normalized at one. The number of industrial workers is denoted by L, and there are L_A agricultural workers. Because changes in

E only arise through changes in the number of (mobile) industrial workers or through changes in their wage rate, we have

$$\begin{aligned} \hat{E} &= \frac{1}{E}(L\mathrm{d}w + w\mathrm{d}L) \\ &= \mu(\hat{w} + \hat{L}) \end{aligned} \tag{2.25}$$

where we have used the fact that in the symmetric equilibrium, the share of manufacturing wages in total output must be equal to the share of manufacturing in consumption, μ.

We totally differentiate the expression for a firm's profit (2.4) to find

$$\begin{aligned} \mathrm{d}\pi &= p\mathrm{d}x - p\theta\mathrm{d}x \\ &= \frac{p}{\sigma}\mathrm{d}x \end{aligned}$$

where we use the definition of θ defined earlier. From the fact that prices p and the elasticity σ are positive, we deduce that changes in profit $\mathrm{d}\pi$ and changes in production $\mathrm{d}x$ have the same sign. This is intuitive: a firm, at equilibrium, produces and sells just enough to cover its total costs. The price has been set higher than marginal costs in this monopolistic competition framework, so any increase in sales beyond the equilibrium quanitity will make the firm profitable. We will look at the sign of $\mathrm{d}x/\mathrm{d}n$ as an indication of the sign of $\mathrm{d}\pi/\mathrm{d}n$. The latter derivative determines the stability of the symmetric equilibrium: if an increase in the number of firms in a region leads to lower profits, some firms will exit and the equilibrium will be restored. This is the case if $\mathrm{d}\pi/\mathrm{d}n$ is negative. For positive values of the derivative, a small change in the number of firms will make the other firms more profitable, causing more entry and a runaway process of agglomeration.

One point is worth expanding upon before we embark on our calculations. We can check the stability of the equilibrium quite easily in the current setup, as 'a small perturbation in the number of firms' is well defined in this case. As there is only one type of industrial firm, any change in its number must take place along the same dimension. In the next chapter, we will introduce several types of firms, which are active in different sectors. In that model, a change in the number of firms in one sector will affect the profitability of the other firms in that sector, but also the profitability of firms in the other sectors. Also, a perturbation in the number of firms can take place in any of the sectors, or in multiple sectors at once, and to find out about the stability of an equilibrium we will have to check all possible perturbations and their associated changes in profits. We will develop a method of doing this efficiently in section 4.2.

We start with formula (R2) from the previous section. Replacing $\hat{p}$ with $\hat{w}$ and $\hat{L}$ with $\hat{n}$, using (2.25) and the relation $\hat{p} = \hat{w} = \mu\hat{q}$, which follows from real wage equalization, we have that

$$\begin{aligned} \hat{x} &= -\sigma\hat{w} + Z\left(\mu(\hat{w}+\hat{n}) + (\sigma-1)\frac{\hat{w}}{\mu}\right) \\ &= \left(Z\mu - \sigma + \frac{Z}{\mu}(\sigma-1)\right)\hat{w} + Z\mu\hat{n} \end{aligned} \tag{2.26}$$

We can eliminate the $\hat{w}$ from this expression by using formula (R1) from the previous section. Once again using $\hat{p} = \hat{w} = \mu\hat{q}$, it can be written as

$$\hat{w} = \frac{-\mu Z}{(\sigma-1)(1-Z\mu)}\hat{n}$$

Using this value for $\hat{w}$ in (2.26), we find

$$\hat{x} = \frac{(2\sigma-1)\mu - [\sigma(1+\mu^2)-1]Z}{(\sigma-1)(1-Z\mu)}Z\hat{n}$$

The sign of $\hat{x}/\hat{n}$ determines the stability of the symmetric equilibrium. The definition of Z from the previous section renders a value of τ that lies on the border between stability and instability. That value is

$$\tau_{\text{break}} = \left[\frac{(\sigma(1+\mu)-1)(1+\mu)}{(\sigma(1-\mu)-1)(1-\mu)}\right]^{\frac{1}{\sigma-1}}.$$

We look next at the sustain point, the value of transport costs at which the asymmetric equilibrium in which all firms have agglomerated in one region, is only just stable. Once again, we use the special form of the equilibrium (total agglomeration) to simplify certain relations in the model, and consider the fate of a breakaway firm. Meanwhile, we retain the assumption that firms enter and exit much slower than workers switch regions. This ensures the equality of real wages in both regions at all times.

When all firms have agglomerated in the *home* region, we know that the expenditures of each region, equal to the total wages paid to its inhabitants, are related by

$$E^* = \frac{1-\mu}{1+\mu}E \tag{2.27}$$

Also, we can simplify the relationship between the two price indices to

$$q^* = q/\tau. \tag{2.28}$$

This is a simplification of formula (2.9) above.

We now look at the demand that a firm receives when it is part of the agglomerated region, and compare it to the demand that it would receive if the firm decided to break away from the agglomeration and move to the peripheral region. For the first case, we add demand from the home region (formula 2.2) and from the foreign region (formula 2.8) to find the demand that a firm in the agglomerated region can expect:

$$x_{\text{core}} = \mu\left[Eq^{\sigma-1} + E^*(q^*)^{\sigma-1}\tau^{\sigma-1}\right]p^{-\sigma} \tag{2.29}$$

We can use the simplifications in (2.27) and (2.28) to write this as

$$x_{\text{core}} = \frac{2\mu}{1+\mu}Eq^{\sigma-1}p^{-\sigma} \tag{2.30}$$

This is the demand that a representative firm faces when it is located in the core, the agglomerated *home* region. For the demand that the same firm would get if it were to move to the *foreign* region, we use a version of (2.29) where the transport costs work the other way:

$$\begin{aligned} x_{\text{periphery}} &= \mu\left[Eq^{\sigma-1}\tau^{\sigma-1} + E^*(q^*)^{\sigma-1}\right](p^*)^{-\sigma} \\ &= \mu q^{\sigma-1}E\left[\tau^{\sigma-1} + \tau^{1-\sigma}\frac{1-\mu}{1+\mu}\right](p^*)^{-\sigma} \end{aligned} \tag{2.31}$$

where we used the same simplifications. We now turn our attention to the ratio $x_{\text{core}}/x_{\text{periphery}}$. This ratio tells us if the agglomerated equilibrium is stable, for if it is smaller than one, a firm can expect more demand in the peripheral region than in the agglomerated region. As above, profits are proportional to demand, so when the ratio is below one we know that profits must be larger in the *foreign* region than they are in the *home* region. This means that the agglomerated equilibrium is unstable. Reversely, a value of the ratio larger than one indicates that the agglomeration is stable. We write

$$\begin{aligned} \frac{x_{\text{core}}}{x_{\text{periphery}}} &= \left(\frac{p}{p^*}\right)^{-\sigma}\left[(1+\mu)\tau^{\sigma-1} + (1-\mu)\tau^{1-\sigma}\right]/2 \\ &= \tau^{\sigma\mu}\left[(1+\mu)\tau^{\sigma-1} + (1-\mu)\tau^{1-\sigma}\right]/2. \end{aligned} \tag{2.32}$$

From this expression, we can calculate the value of τ for which the agglomerated equilibrium is just stable: in that case, the ratio is equal to one. This holds trivially for $\tau = 1$; if there are no transport costs, location is irrelevant and firms will receive equal demand wherever they locate. If we differentiate (2.32) with respect to τ at $\tau = 1$, we find that the derivative is equal to $(3\sigma - 2)(\mu - 1)$, which is negative. This means that a small decrease in τ (or, the introduction of transport costs) leads to a situation where the ratio in (2.32) becomes larger than one. In that case, agglomeration is stable.

Neary (2001) shows that, for most practical values of μ, there exists a second, fractional value of τ for which the expression in (2.32) is equal to one. He also proves that it must be higher than τ_{break} derived above, in which case there exist values of transport costs for which the agglomerated equilibrium is stable, as is the symmetric equilibrium. In that case, history decides which equilibrium obtains.

Chapter 3

A model with continuous sectors

During the years, an odd trend may be observed in the economic geography models that were proposed. In the early papers, such as Fujita (1988) and Rivera-Batiz (1988), the space in which the location problem was solved was a one-dimensional line. The position on this line captures the distance from a central business district, and once the problem is solved in one dimension a two-dimensional solution is trivially available (This line of modelling goes back to von Thünen 1842). In later papers, starting from Krugman (1991a), the spatial structure was simplified to two possible locations with fixed transport costs between them. The internal ordering of the locations was left undiscussed. So, in some ways, the geographical predictions of new economic geography turned out to be weak.

One particular way of modelling linkages is used in Krugman and Venables (1995). In this paper, agglomeration occurs because of input-output connections: the production of one firm is used as an input for other firms, who therefore prefer to be close by. The authors assume that every firm uses as an input a composite good, made from the output of all other firms. In fact, the structure of production is completely symmetric, so that each firm uses every possible product to the same degree. This is a natural property of the Dixit-Stiglitz framework, in the way it is usually applied.

It is not impossible to step away from the assumed symmetry and allow for some variety in the input structure. Young (1993) introduces a production function with a continuum of intermediate goods suppliers and a continuum of final goods suppliers. The latter use intermediate goods, but only if these goods come from producers that are sufficiently 'close' to their product, where closeness is defined within the continuum of firms.

In Section 3.1, I introduce a similar adaptation of the Dixit-Stiglitz framework. On top of this adaptation, we can construct a model of economic geography à la Krugman and Venables (1995). For a few simple input-output patterns, this model shows that a greater variety of equilibria may be obtained than just agglomeration or symmetry. We develop and experiment wit this model in Section 3.2. We look at a slightly more complicated form of the input-requirement function, which defines our sectors, in Section 3.3. In this section, we prove the existence of an equilibrium in which 'old' firms cluster in one region, while 'new' firms agglommerate in the other. This result allows us to characterize the development of different regions in a growing economy. Section 3.4 concludes.

3.1 Generalized Monopolistic Competition

We look at production that takes place in firms that operate in a monopolistic competitive[1] market. As inputs, these firms use labor and composite of output of other firms. The production function is of the Cobb-Douglas variety,

$$z_i = L_i^{\alpha} Q_i^{1-\alpha}. \tag{3.1}$$

The set of firms is assumed to be a continuum $[0, n]$ [2] Each firm produces a single product that may likewise be indexed on $[0, n]$. We can now define the composite Q_i, that firm i uses as an input, as a CES-aggregate of those products.

$$Q_i = \left[\int\limits_{f(i)} \left(x_i^j \right)^{\theta} \mathrm{d}j \right]^{1/\theta} \tag{3.2}$$

where x_i^j is the amount of input j used by firm i and $0 < \theta < 1$. Here, we use the function $f(i) : [0, n] \to S_n$, where S_n is the sigma field of all open and closed intervals on $[0, n]$. This function indicates what array of inputs firm i uses, and we require nothing of it except that it does not map to the empty set so that each firm uses at least a positive measure of intermediate inputs.

In the usual application, $f(i)$ is the entire set $[0, n]$ for all i. This accounts for the symmetry in production structure that we discussed in the introduction. Young (1993) uses the function $f(i) = [Bi, \min(\Theta i, n)]$ where $0 \leq B < 1$ and $\Theta > 1$.

Because of the continuum of firms that we assumed, this modification of Dixit and Stiglitz (1977) does not alter the market structure very much. Firms still face an MC market where their pricing decision does not affect the general price level. More specifically, a producer i who spends an

---

[1] We will use the acronym MC from now on.

[2] For an interpretation of such a continuum of firms, see section 2.A

amount E on intermediate goods will demand of good j the amount x_i^j:

$$\begin{aligned} x_i^j &= \frac{Ep\left(j\right)^{-\sigma}}{\int\limits_{f(i)} p\left(k\right)^{1-\sigma} \mathrm{d}k} \text{ if } j \in f\left(i\right) \\ &= 0 \text{ if } j \notin f\left(i\right). \end{aligned}$$

where $\sigma = 1/\left(1-\theta\right) > 1$. This shows that producers are confronted with a constant price-elasticity of demand σ, when the demand is for intermediate goods. If final demand for the goods also has the same elasticity σ, the optimal pricing decision is to set prices as a markup over marginal costs. The optimal markup is $\sigma/(\sigma-1)$.

The price index for producer j's intermediate good is

$$p_Q^j = \left[\int\limits_{f(j)} p\left(k\right)^{1-\sigma} \mathrm{d}k\right]^{\frac{1}{1-\sigma}}. \tag{3.3}$$

This is the 'ideal' price index (see Green 1964) so that $p_Q^j Q_j = E$.

Based on the intermediate-input function $f(i)$ we can define an 'inverse' function $g(i)$ that maps the index of a producer, i, into the set of intermediate producers that use her good:

$$\begin{aligned} g(i) &: [0,n] \to S_n \\ g(i) &= \{j \in [0,n] | i \in f(j)\}. \end{aligned}$$

We assume that all intermediate goods are used somewhere, so that the function $g(j)$ does not map to the empty set for any j in $[0,n]$. Using this function, we can write the demand for a specific intermediate good j as

$$\begin{aligned} x^j &= \int\limits_{g(j)} \frac{E(i)p\left(j\right)^{-\sigma}}{\int\limits_{f(i)} p\left(k\right)^{1-\sigma} \mathrm{d}k} \mathrm{d}i \\ &= p(j)^{-\sigma} \int\limits_{g(j)} \frac{E(i)}{\int\limits_{f(i)} p\left(k\right)^{1-\sigma} \mathrm{d}k} \mathrm{d}i \end{aligned}$$

Notice that producer j still faces a demand curve with constant elasticity σ, as in the MC setup. The price that maximizes profit will therefore be a markup $\sigma/\left(\sigma-1\right)$ times marginal cost, as usual.

3.1.1 An example

To gain some insight into the effects of this modification, let us look at two examples that will prove useful later on. We first model a recursive[3] MC

---

[3] The term 'recursive' is used to indicate that the group of producers uses some of its own product as input (see eqn. 3.1)

economy where there is no variety in input structure, and then one where there are two distinct sectors. The latter example can be generalized to N distinct sectors.

Example 1: A one-sector economy

We consider an economy with L workers that supply one unit of labor inelastically. We use the wage rate as numéraire and set it to one. There is a continuum $[0, n]$ of firms and n is fixed. Producers face production functions (3.1) and (3.2). In this example, we take $f(i) = [0, n]$ for all i. This implies that a firm's cost function is

$$\begin{aligned} C\left(p_Q, z_i\right) &= (1-\alpha)^{\alpha-1}\alpha^{-\alpha} w^{\alpha} p_Q^{1-\alpha} z_i \\ &= (1-\alpha)^{\alpha-1}\alpha^{-\alpha} p_Q^{1-\alpha} z_i \end{aligned} \tag{3.4}$$

The price index p_Q is defined in (3.3); the superscript i is omitted because it is the same for all firms, due to our assumption about $f(i)$.

Consumers maximize utility U, given by

$$U = \int_0^n \left(x_i^j\right)^{\theta} \mathrm{d}j. \tag{3.5}$$

Notice the crucial assumption that the parameter θ is the same in this utility function and in production function (3.2). A producer, whose product is demanded from consumers as well as other producers, now faces two demand curves with the same constant elasticity σ. This reduces the problem, as the optimal price is simply a markup $\sigma/(\sigma-1)$ times the marginal cost, or

$$p = \frac{\sigma}{\sigma-1}(1-\alpha)^{\alpha-1}\alpha^{-\alpha} p_Q^{1-\alpha}$$

from (3.4). Because p_Q is again a function of p as in (3.3), we can simplify to

$$p = \left(\frac{\sigma}{\sigma-1}\right)^{\frac{1}{\alpha}} (1-\alpha)^{\frac{\alpha-1}{\alpha}} \alpha^{-1} n^{\frac{1-\alpha}{\alpha(1-\sigma)}} \tag{3.6}$$

where we have used the fact that all prices are equal in equilibrium. Notice that this formula completely fixes the price in terms of parameters. We can now also solve for the other endogenous variables.

The price index for the composite good Q is, by (3.3), equal to

$$p_Q = p n^{\frac{1}{1-\sigma}}.$$

Because labor supply is exhausted, we know that each 'unit of firms' applies L/n units of labor in equilibrium, which is also the sum it pays out in wages. From the Cobb-Douglas structure of (3.1), we know that this must be a fraction α of total costs. Therefore,

$$Q_i = \frac{L}{n}\frac{1-\alpha}{\alpha}\frac{1}{p_Q} \tag{3.7}$$

for all i. This is the amount of the composite good that firm i demands. From each specific producer j, an amount x_i^j is demanded, where

$$\begin{aligned} x_i^j &= Q_i p(j)^{-\sigma}\left[\int_0^n p(k)^{1-\sigma}\mathrm{d}k\right]^{\frac{\sigma}{1-\sigma}} \\ &= Q_i n^{\frac{\sigma}{1-\sigma}}, \end{aligned}$$

where we used the fact that all prices are equal in equilibrium. This means that each firm j faces an intermediate demand for its good equal to

$$\int_0^n x_i^j \mathrm{d}i = Q_i n^{\frac{1}{1-\sigma}}. \tag{3.8}$$

We now look at final demand. Because firms use a constant returns technology and price as monopolists, they will earn a nonzero profit. We assume that each inhabitant owns an equal stake in each firm, so that these profits are equally divided among them. This puts per capita income at

$$y = 1 + \pi\frac{n}{L} \tag{3.9}$$

where π is the profit of a firm. From the markup pricing rule, we know that profits are a fraction $1/\sigma$ of wholesale, so that

$$\pi = \frac{1}{\sigma}z_j p. \tag{3.10}$$

The total expenditure on all goods then is Ly, so allocated that (3.5) is maximized. This gives a per-firm final demand of

$$y_j = L\frac{y}{p_Q}n^{\frac{\sigma}{1-\sigma}} \tag{3.11}$$

As a check on our computations, it is now possible to show that final plus intermediate demand per firm is equal to a firm's production, or

$$L\frac{y}{p_Q}n^{\frac{\sigma}{1-\sigma}} + Q_i n^{\frac{1}{1-\sigma}} = \left(\frac{L}{n}\right)^{\alpha} Q^{1-\alpha}. \tag{3.12}$$

This is done in Appendix 3.A.

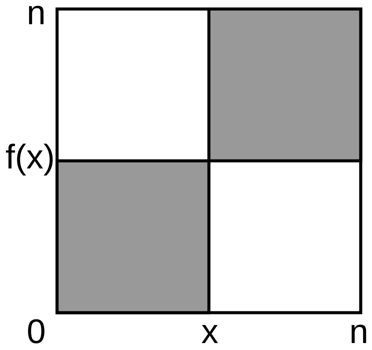

Figure 3.1: Function $f(x)$ in formula (3.13) mapped out.

Example 2: Two distinct sectors

We use the same setup as in the previous paragraph, with the exception that we alter $f(x)$ and use

$$f(x) = \left\{ \begin{array}{ll} [0, \frac{1}{2}n) & \text{if} \quad x \in [0, \frac{1}{2}n) \\ [\frac{1}{2}n, n] & \text{if} \quad x \in [\frac{1}{2}n, n] \end{array} \right. \tag{3.13}$$

With this f, there are two sectors in the economy, separated at $x = n/2$ in the continuum. Firms only use intermediates from their own sector, and consequently only receive demand for intermediates from their own sector. In this specific case, we have $f(x) = g(x)$ for all x.

It is possible to visualize $f(x)$, as is done in Figure 3.1. The index x is on the horizontal axis, while $f(x)$ may be read of the vertical axis as the grey area above x. Alternatively, the function $g(x)$ could be read of the horizontal axis with x on the vertical. In this case, both functions are the same.

We solve the problem along the same lines as in section 26. As a consequence of the change we made, we must study some of the characteristics of the two sectors separately. However, because the sectors are exactly the same, we can suffice with the specification of one of them; the same results will hold for the other.

Within the first sector, $[0, n/2)$, the price index for the composite good is different from above, and now reads

$$p_Q^j = \left[\int_0^{\frac{1}{2}n} p(k)^{1-\sigma} \mathrm{d}k \right]^{\frac{1}{1-\sigma}}. \tag{3.14}$$

Therefore, the price in sector 1 will be

$$p = \left(\frac{\sigma}{\sigma - 1} \right)^{\frac{1}{\alpha}} (1 - \alpha)^{\frac{\alpha - 1}{\alpha}} \alpha^{-1} \left(\frac{1}{2} n \right)^{\frac{1-\alpha}{\alpha(1-\sigma)}} \tag{3.15}$$

Note that this is formula (3.6) multiplied by $(1/2)^{(1-\alpha)/(\alpha[1-\sigma])}$. Because the last exponent of this expression is negative, we see that this price in (3.15) is higher than the price in (3.6). This reflects the restriction that we have imposed on sector 1 firms: they can only use inputs from a subset of firms. Because there are increasing returns to the scale of available inputs (see chapter 2), a decrease in scale will increase the costs and therefore the price.

From (3.14) we have that $p_Q^j = p(\frac{1}{2}n)^{\frac{1}{1-\sigma}}$, which is used in the unchanged formula for Q_j, formula (3.7). Knowing Q_j, we know how much each firm in the sector spends on the intermediate composite good. For every single firm, this means a demand of $Q_j(\frac{1}{2}n)^{\sigma/(1-\sigma)}$ from every firm in its sector, leading to a per-firm demand for intermediate purposes of

$$Q_j(\frac{1}{2}n)^{\sigma/(1-\sigma)} \cdot \frac{1}{2}n = \frac{1}{2}L\frac{1-\alpha}{\alpha}\frac{1}{p_Q^j}\left(\frac{1}{2}n\right)^{\frac{\sigma}{1-\sigma}}$$

As for final demand, the public still consumes *all* goods and maximizes (3.5). Therefore, the price index faced by the public still is

$$p_Q = pn^{\frac{1}{1-\sigma}} \tag{3.16}$$

with p from (3.15). This index is used in (3.11), while formulae (3.9) and (3.10) still hold in this model.

Example 3: N distinct sectors

We can generalize the above case further by taking the number of sectors a variable N and using the function

$$f(x) = \begin{cases} [0, \frac{1}{N}n) & \text{if} \quad x \in [0, \frac{1}{N}n) \\ [\frac{1}{N}n, \frac{2}{N}n) & \text{if} \quad x \in [\frac{1}{N}n, \frac{2}{N}n) \\ \vdots & \vdots \qquad \vdots \\ [\frac{N-1}{N}n, n] & \text{if} \quad x \in [\frac{N-1}{N}n, n] \end{cases} \tag{3.17}$$

A map of the function is in Figure 3.2 for $N = 5$. Again we assume that firms in a sector only use products from their own sector.

Instead of solving the entire model again, this time we make use of an interesting regularity in the outcomes. From the inelastic supply of labor, the fact that wage is fixed at 1 and the number of firms is fixed, we know that the wage bill per firm is always the same, L/n. Due to the Cobb-Douglas structure of function (3.1) this also fixes the amount of money spent on intermediates. From that amount, it is possible to compute that a constant fraction ϕ of the production of any firm is used as intermediates by other firms in the same sector. there holds that

$$\phi = (1-\alpha)\,\frac{\sigma-1}{\sigma},$$

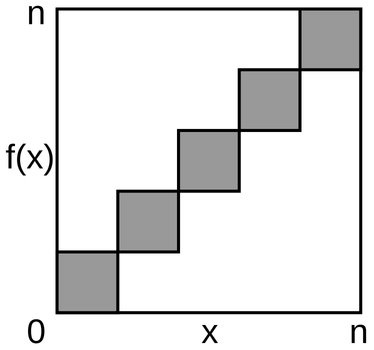

Figure 3.2: Function $f(x)$ in formula (3.17) mapped out.

so that ϕ independent of N.

The number ϕ can help compute output quickly. This works as follows: we know that the size of a sector is n/N, which determines the intermediate market, and is also the length over which the integral and p_Q^j is computed. Knowing this size, we can compute p as in (3.6) and (3.15). From p and the size, we compute p_Q^j, and from this we know Q from (3.7), which in turn fixes intermediate demand. And with that in hand, we can use ϕ to determine total production per firm.

It turns out that production per firm is

$$z_j = \phi^{\frac{1-\alpha}{\alpha}} \frac{L}{n} \left(\frac{N}{n} \right)^{\frac{1-\alpha}{\alpha(\sigma-1)}} .$$

In this expression we can clearly see the effect of dividing up the economy into different sectors. The exponent on N is negative, so that the more sectors there are, the less is produced in total. This is caused by the fact that there are increasing returns to the number of varieties available (or, usable) as intermediate input.

It is important to bear in mind that the above result hinges on the fact that there is no inter-sector trade. This assumption is not realistic, but can easily be relaxed.

3.2 A two-sector model of economic geography

In this section, we will use the model of example 2, section 3.1.1, in an environment with two locations and nonzero transport costs. This setup is reminiscent of Krugman and Venables (1995), albeit that the economy now has two, separate, sectors.

3.2.1 Sectors that only use their own product

Imagine there are two possible locations, N and S. Transport costs are of the iceberg-kind as in Samuelson (1952): only a fraction τ of the goods that are shipped actually arrives. Each location has population $L/2$, which is the amount of labor supplied with elasticity zero. The wage rate in N is one by normalization, that in S is w.

We again assume a continuum of firms with length n. Of these firms, those in $[0, n/2)$ are in sector 1 and those in $[n/2, n]$ are in sector 2. The measure of firms per sector is invariant. We denote by $m_1 \in [0, n/2]$ the measure of firms from sector 1 that reside in location N. That leaves $n/2 - m_1$ firms from sector 1 in location S. Similarly, $m_2 \in [0, n/2]$ firms from sector 2 reside in location N.

Firms use labor and an intermediate composite good, which comprises output from all the firms in their own sector. The price index for the composite good in section 3.1.1 was given by (3.14). Now that there are transport costs, this index is a little more complicated. For a firm in sector j and location λ, it is

$$p_Q^{j,\lambda} = \left[\int_0^{n/2} \left(\frac{p(k)}{\tau^{|\lambda - L(k)|}} \right)^{1-\sigma} \mathrm{d}k \right]^{\frac{1}{1-\sigma}} \tag{3.18}$$

where $j \in \{1, 2\}$ is the sector and L indicates the location of firm k:

$$L(k) = \begin{cases} 0 & \text{if firm } k \text{ is in } N \\ 1 & \text{if firm } k \text{ is in } S \end{cases} .$$

The same convention holds for the variable λ.

As above, we assume that final demand is for a composite of all goods from both sectors. The price index for that composite, previously given by (3.16), now is different for the two locations, and equal to (3.18) with the integral along $[0, n]$.

Producers face a demand curve that is an aggregate of demand from firms in two regions and consumers in two regions. Because the costs of transport are just a multiple of wholesale, and because consumers and firms share parameter θ, this curve has a constant elasticity of demand σ. Optimal prices are therefore a markup $\sigma/(\sigma - 1)$ over marginal cost.

The solution to this model may now be derived. All price indices take the form

$$\left(\sum_{i=1}^{l} \Phi_i \right)^{1/(1-\sigma)}$$

with l an integer larger than one. Contrary to section 3.1, the terms Φ_i now vary with i. This form cannot be simplified, so that we must rely on computational methods to approximate a solution. This necessity almost

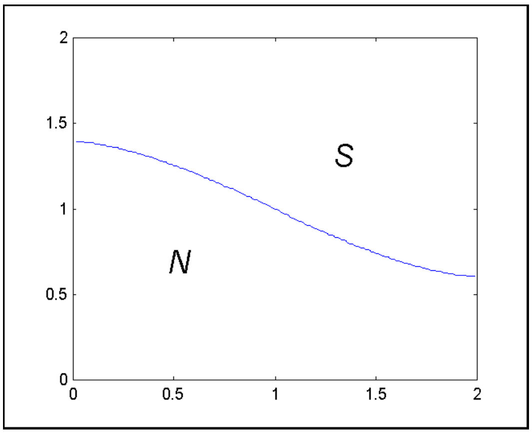

Figure 3.3: Sector 1 preferable region

always arises in new-geography models, and is recognized by Krugman (1998).

The approximation works as follows. Given the wage, prices can be computed. Given the amount of labor hired by each firm (which must integrate to $L/2$ for each location), production and demand can be computed.

With those results, we can look at how the parameters that were taken as given should be modified. Labor hired reacts to excess demand or supply of goods in a sector-location couple. Wage responds to excess demand or supply between the two locations.[4] The model converges until all demand, intermediate and final, is equal to supply.

With the above solution, we have taken the measure of firms per location, m_1 and m_2, as given. One of the results from the model is that we can compute the profit per firm, as a function of location and sector. This is because all firms in the same sector and in the same location behave alike, and have the same profit. Looking at the pattern of profits can give some insight into the possible migration patterns of firms, assuming that they are driven by profit maximization. Note that a migrating firm leaves its laborers behind and hires from the other pool, so that the number of inhabitants remains equal between the locations.

The results are in the two figures above. The variable m_1 is on the horizontal axis, m_2 is on the vertical axis. The left panel shows the preferred location for firms in sector 1, given m_1 and m_2. The preferred location is the location where the profits per firm are higher. The same diagram is drawn in the right panel, for sector 2. The other parameters in this model were $n = 4$, $L = 40$, $\sigma = 3$, $\tau = 0.8$ and $\alpha = 0.6$.

To find the agglomeration pattern that might result if the firms actually

---

[4] Instead of varying the wage we could also have specified an exchange rate between the two locations and set both wage rates to one—in their own currency.

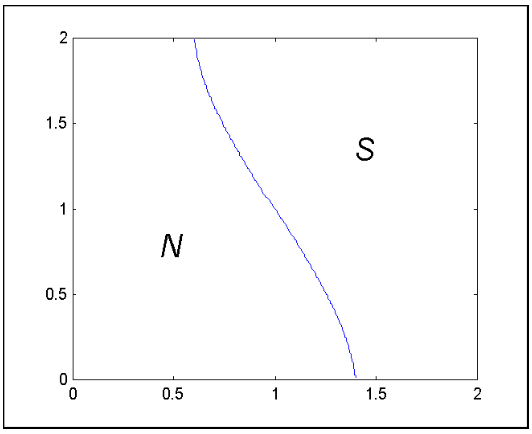

Figure 3.4: Sector 2 preferable region

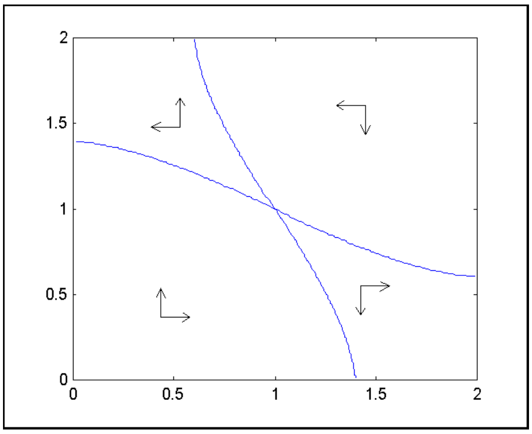

Figure 3.5: Direction of motion in the (m_1, m_2) plane. m_1 on the horizontal axis, m_2 vertical.

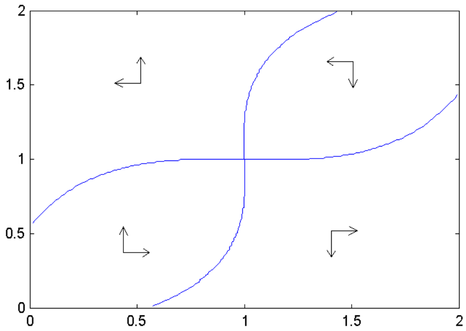

Figure 3.6: Dynamics when $\tau = 0.2$

responded to the incentives given by the profit rate, we combine the two panels in figure 3.5. This figure shows that the model, with the current parameters, tends to correct imbalances. If there are few firms of both sectors in region N (low m_1, m_2), there will be migration toward that region. However, if the imbalance is such that there are a lot of firms of one sector in the north, while most of the firms of the other sector are in the south, the tendency is toward complete separation of the two sectors. There are three long-term equilibria in this model: the saddle-point stable equilibrium $(m_1, m_2) = (n/4, n/4)$ and the stable equilibria $(0, n/2)$ and $(n/2, 0)$.

The precise long-term result depends on how the laws of motion of the firms are specified, and on the initial condition. If region N historically has a lot of sector 1 activity, while region S is the historic center for sector 2, we see that the model reinforces that structure.

This result is interesting because it is reminiscent of many other results in economic geography. By that I mean the dependence on initial conditions and complete agglomeration of sectors. However, the division of the economy into sectors adds to the credibility of the outcome. No longer does *all* activity agglomerate into one location, as previous results showed, but we have a situation where the agglomeration is per sector. This is because the economies of scale that drive agglomeration are present within a sector, but the diseconomies of scale (*e.g.*, rising wages) are present between sectors.

It can be interesting to modify the parameters a bit and to check the effects on the outcome. In figure 3.6, we increased transport costs tremendously by setting $\tau = 0.2$. It turns out that there still are three equilibria with the same stability properties. This is typical for all values of $\tau < 1$.

3.2.2 Other IO patterns

We look at a number of other patterns of input-output between sectors, using the same methodology as above. First, consider the two-sector situation

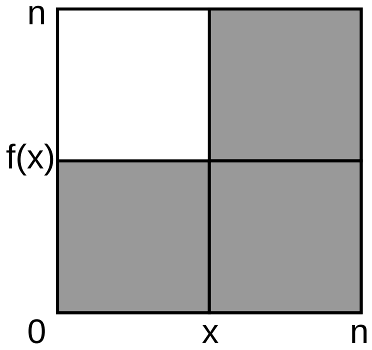

Figure 3.7: The function $f(x)$ in formula (3.19) mapped out

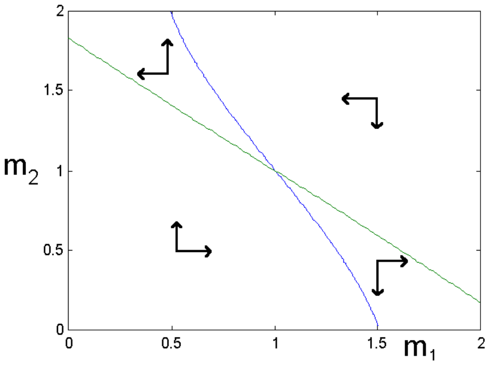

Figure 3.8: Direction of motion when sector 1 uses all output as intermediate, sector 2 uses only sector 2 output.

where sector one uses a composite of *all* output as intermediate and sector two uses only its own output. The function f in this case is

$$f(x) = \begin{cases} [0, n] & \text{if } 0 \leq x \leq n/2 \\ [n/2, n] & \text{if } n/2 < x \leq n \end{cases} \tag{3.19}$$

In this case, if there is a region with many sector two firms, the firms from sector one face conflicting incentives. Because of the 'crowding out' phenomenon, wages will be relatively high in that region, so that it is relatively unattractive. On the other hand, the new input-output structure shows that there is an advantage in being close to sector two firms, as that is where sector one firms get some of their inputs from. The contradictory incentives are clear from figure 3.8, with the directions of motion.

We see that for high values of m_2, and low values of m_1, there is almost always an incentive to increase m_1. Compared with figure 3.6, the two indifference curves have moved towards each other, leaving a smaller

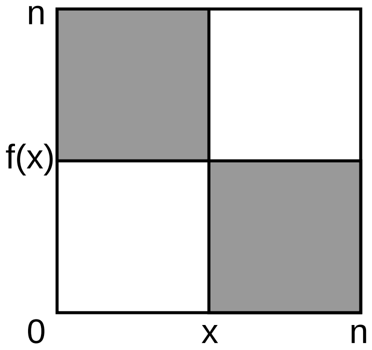

Figure 3.9: The function $f(x)$ in formula (3.20) mapped out.

area with divergent behavior. Figure 3.8 uses the same parameters as the simulation in figure 3.6 above.

It is not surprising that for sector two firms, the crowding-out effect is also less. The presence of a large number of sector one firms increases the demand for their product, which makes the location more attractive. However, we see that of the three equilibria $(2,0)$, $(0,2)$ and $(1,1)$, still only the first two are stable. So, we still have a model where the long-term solution is that each sector finds its own region.

We can also do a complete reversal of the model in Section 3.2.1 and adapt the input-output function

$$f(x) = \begin{cases} [n/2, n] & \text{if } 0 \leq x \leq n/2 \\ [0, n/2] & \text{if } n/2 < x \leq n \end{cases} \tag{3.20}$$

This function specifies that firms in sector one use a composite of sector two output as input, and vice versa. The sectors do not use any of their own output as intermediate.

The results are in Figure 3.10. In this case, there is no advantage for firms in being isolated with their own kind. There is only one equilibrium, and it is the stable equilibrium $(1,1)$.

3.3 Location, sectors and economic growth

3.3.1 Introduction

In the previous paragraph, we have shown how it can be in the interest of industries from different sectors to settle in different regions. The strength of the ties within and between those sectors, which is measured by the input-output function, determines the equilibrium location outcome. Which industry settles in one particular region is indeterminate, but once

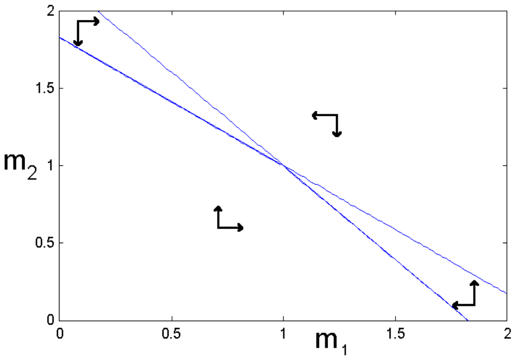

Figure 3.10: The direction of motion when sectors use each others output as intermediate input.

the equilibrium is attained it may self-perpetuate. If the ties between sectors are such that firms like to settle close to other members of the same sector, as in section 3.2.1, the initial characteristics of a region are locked in when new firms arrive. The new firms will locate close to members of their own sector.

This outcome would allow us to make certain observations on the development of regions, if we are prepared to assume something about the ties that new firms will have with different existing sectors in the economy. We will make such an assumption in this section. Specifically, we will assume that ties between firms are stronger if they are approximately from the same period. We will propose a specific form of the IO-function $f(x)$ that reflects this assumption and use it to derive a spatial equilibrium between two regions. The specific form of $f(x)$ will then allow us to discuss the evolution of this equilibrium in the context of a growing economy.

Throughout this chapter, we have assumed that the number of firms n is fixed, indicating that there is no free entry for firms. We now expound on this assumption: in our model, specific technical knowledge is required to start a firm. This knowledge is proprietary and can only be used to start a single firm. The assumption allows the owners of the firm to turn a profit, something that is impossible under free entry. Up to now, we have kept looked at a situation in which the number of firms is constant; in this situation, no new technical developments take place.

The economy as we observe it, with a continuum of n active firms, is the result of technological progress from the past. We will assume the following about the nature of this progression: each new production process is partly the result of insights gleaned from a number of recent previous innovations, and uses the products of those innovations as inputs. In turn, a new production process can itself inspire new products. Those new products can then serve as an input into the original process, making it more

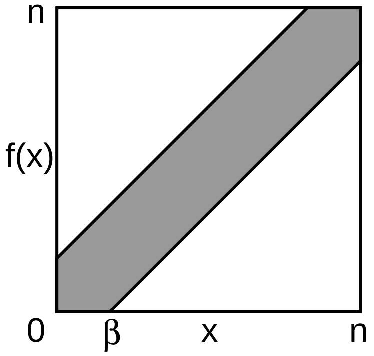

Figure 3.11: Function $f(x)$, which shows which intermediate goods each firm $[0 \ldots n]$ uses, as defined in formula (3.21).

efficient. In this setup, each firm uses as intermediate inputs those products that are 'close' to its own production process in a technical sense. This assumption was used by Young (1993), for instance, to explain certain aspects of economic growth.

In a sense, our previous example 2 on page 60 is a crude representation of this assumption. Firms in one sector can be thought to be technologically close to each other. If we view the order of firms on the line $[0 \ldots n]$ as the result of a technological progression, a new technological paradigm was introduced in firm $n/2$ which resulted in incompatibility between the two sectors.

We will abstain from such shifts in paradigm in our example and use a continuous mapping $f(x)$, defining a gradual scale of technological processes. We use

$$f(x) = [\max(x - \beta, 0), \min(x + \beta, n)] \tag{3.21}$$

where $\beta > 0$. Again, $f(x)$ is the range of firms from which firm x uses the output as an intermediate input. A graphical representation of this function is given in figure 3.11. The parameter β is a measure of how 'close' firms have to be to another firm, in order use their product as an intermediate input. For large values of this parameter, if $\beta > n$, we are back at the first model of this chapter where each firm uses every available product as an intermediate input. For smaller β's, firms only use the products of their technological neighbors. Note that for firms close to n, the set of available intermediate goods is smaller than for those in the middle. This is borne out by the minimum-condition in formula (3.21). These firms would like to make use of more advanced intermediate products, made by firms with an index larger than n. These, more advanced, products have not yet been developed.

We will discuss a particular spatial equilibrium for an economy in which this function $f(x)$ determines the input-output structure. In our discussion, we will leave the value of β unspecified. This will allow us to make a few comments about regional specialization and economic growth afterwards. Economic growth, driven by technological innovations, can be seen in this model as an increase in n, the number of firms. However, an increase in n is isomorph with a *decrease* in β. Because our equilibrium holds for all values of β, it is not affected by economic growth. We will return to this subject after our discussion of the spatial equilibrium.

3.3.2 The spatial equilibrium

We use the model of the previous sections, including the spatial setup from paragraph 3.2: there are two regions with transport costs between them, so that only a fraction $\tau < 1$ of transported goods arrives. For a function that determines the nature of the intermediate goods aggregate for each firm x, we use the specification in (3.21).

In this paragraph, we will prove that the symmetric equilibrium of this model is stable. The symmetric equilibrium is given by the following solution to the location problem: firms with an index between 0 and $n/2$ locate in one region, firms with an index between $n/2$ and n locate in the other. Throughout the section, we will assume that firms with a lower index locate in the N region. This assumption is, of course, immaterial.

We prove that the symmetric equilibrium is stable by showing that, given that the equilibrium has obtained, no firm would want to deviate from it. This does not prove that other types of equilibria are impossible; we simply show the existence of this particular equilibrium. The fact that the solution is symmetric simplifies the proof considerably: apart from the numbering of the firms, the situation of firm $i < n/2$ in region N is exactly identical to that of firm $n - i$ in region S. One result of this symmetry is that wages in both regions are the same. We will exploit the symmetry of the equilibrium throughout the proof.

Consider firm i, with $i < n/2$, in region N. We will show that this firm does not want to move from its home region to region S. Because of the requirements to start a firm (specific proprietary technical knowledge) it is impossible to copy a version of this firm to the other region, so that the decision not to move is leaves the equilibrium in place. As stated above, the economic circumstances of firm i are identical to those of firm $n - i$ in the other region. By proving that firm i does not want to leave, we will have proved the same about firm $n - i$, and thus about all the firms in region S.

The profits of firm i are determined by a number of factors, some of which vary with the region in which the firm operates. These factors are: local wage, final demand, intermediate demand and intermediate costs. Symmetry implies that local wage is the same in the two regions, as is final demand: we retain the assumption that the (identical) consumers of both regions want each available product equally much. The location decision of firm i, which is infinitely small, does not change this. Thus, we only look at intermediate demand and the costs of intermediate goods for firm i in each region. We will show that the costs are lower in region N, and that demand is higher in that region.

First we look at intermediate costs. Consider two possible cases: in the first case, $i + \beta \leq n/2$. This means that all intermediate goods suppliers of firm i are located in the same region, the home region N. It is immediately obvious that relocating to the other region will increase intermediate costs with the transport markup. In the second case, $i + \beta > n/2$. We look at the price index of intermediate goods for firm i, which is similar to formula (3.18). If we use the the same definition of the location function $L(k)$ as on page 63, it is is equal to

$$\begin{aligned}\left(p_Q^{i,N}\right)^{1-\sigma} &= \int_{\max(i-\beta,0)}^{\min(i+\beta,n)} \left(\frac{p(k)}{\tau^{L(k)}}\right)^{1-\sigma} \mathrm{d}k \\ &= \int_{\max(i-\beta,0)}^{\frac{n}{2}} p(k)^{1-\sigma}\mathrm{d}k + \int_{\frac{n}{2}}^{\min(i+\beta,n)} \left(\frac{p(k)}{\tau}\right)^{1-\sigma} \mathrm{d}k\end{aligned}$$

The second step follows from the definition of the equilibrium, which stipulates that firms $[0 \ldots n/2]$ are in region N. The second integral is taken over a positive domain in this case, indicating that some intermediate goods suppliers are in the other region. Moving to this region will decrease the costs of their products, but increase those of the region-N suppliers. We rewrite the expression to find out about the balance between the two:

$$\left(p_Q^{i,N}\right)^{1-\sigma} = \int_{\max(i-\beta,0)}^{\frac{n}{2}} \left[p(k)^{1-\sigma} + \left(\frac{p(k)}{\tau}\right)^{1-\sigma}\right] \mathrm{d}k + \int_{\frac{n}{2}}^{n-i-\beta} p(k)^{1-\sigma}\mathrm{d}k$$

where we have used the fact that, due to symmetry, $p(n-k) = p(k)$. Using similar steps, we can write the costs of intermediate goods for the same firm, if it relocates to the S-region, as

$$\left(p_Q^{i,S}\right)^{1-\sigma} = \int_{\max(i-\beta,0)}^{\frac{n}{2}} \left[p(k)^{1-\sigma} + \left(\frac{p(k)}{\tau}\right)^{1-\sigma}\right] \mathrm{d}k + \int_{\frac{n}{2}}^{n-i-\beta} \left(\frac{p(k)}{\tau}\right)^{1-\sigma} \mathrm{d}k$$

Because of the nonnegative domain of the second integral, the fact that prices are positive and because $\tau < 1$, we see that $p_Q^{i,N} < p_Q^{i,S}$. That is, intermediate goods for firm i are cheaper in the N region than in the S region.

We now turn to the intermediate demand for firm i. As a result of the IO-function $f(x)$ in (3.21), firm i receives intermediate demand from firms between $\max(0, i-\beta)$ and $\min(n, i+\beta)$. We again divide these firms in two groups, those in the home (N) region and those in the foreign region. If all demanding firms are in the home region (when $i + \beta \leq n/2$), then moving to the foreign region will certainly decrease intermediate demand because of transport costs. If not, we can identify a subset of the demanding firms, the group $[n - i - \beta, i + \beta]$. Total intermediate demand from this group for firm i will be exactly the same, regardless of where firm i decides to locate. This follows from the symmetry of the initial equilibrium and the symmetry of this particular group around $n/2$. The rest of the firms that demand intermediate goods from firm i, $[\max(0, i - \beta), n - i - \beta]$ is located in the home region N. This means that their demand will be higher if firm i is also located in this region. Taking the separate conclusions together, we see that intermediate demand for firm i is higher if it decides to locate in the initial region N.

Firm i resides in region N, according to the initial equilibrium. We have shown that moving to the other region would change two things for this firm: intermediate goods become more expensive and intermediate demand drops. This will make a move to the other region unattractive: both effects will serve to decrease its profit. We conclude that the position of the firm in the initial equilibrium is stable. This proofs that the equilibrium itself is stable, as the arguments holds for all firms in N and, by symmetry, for all firms in S.

3.3.3 Growth

Now that we have shown the existence of a stable symmetric equilibrium in the static model, let us turn to the situation in which the economy grows over time. Specifically, what will happen to the spatial equilibrium when new firms enter the market?

Our conclusion will depend on the relationship between the new firms and those already present. Let us assume, as we did before, that the numbering of firms from 0 to n implies something about the order in which they were created. Let firm 0 be the oldest firm in the economy, using ancient technology and intermediates from other venerable firms. On the other side, let firm n be the latest addition to the spectrum of technologies. We further assume that any new firms that enter the economy will use a technology related to that of the most modern firm. We can then model an increase in the number of firms from n to $n+\delta$ as a continuation of the present structure: as illustrated in figure 3.12 below, the new firms are added at the end and do not change the form of the input-output relationship.

Where will these new firms locate, if we start with the symmetric equilibrium of the previous section? Naturally, they would like to be close to their intermediate-goods suppliers, the majority of whom will reside in region S.

What will happen next? For a while, the symmetric equilibrium will be disturbed: there will be more firms in region S than in region N. We have

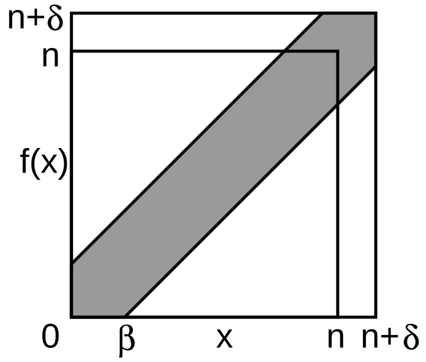

Figure 3.12: Function $f(x)$, defined in formula (3.21), when the number of firms grows from n to $n + \delta$.

seen in previous sections that the mechanism that returns equilibrium is the labor market: increased competition for labor in region S will drive up wages in that region. This will decrease profits for all firms in region S, which leads to the next step in the return to equilibrium: firms will start to think about moving to the N region, where costs are now lower. Which firms will move first? Those with indices close to $n/2$, who already find a large part of their suppliers in the other region, will be the first to find it profitable to go.

When the spatial equilibrium starts changing, where will it end? Without numerical simulations and an explicit rule for relocating between regions we cannot be absolutely sure, but we do know that there is a new symmetric equilibrium close by. Observe that the new input-output function in figure 3.12 is isomorphic to the old function, in figure 3.11 on page 70 above. That is, apart from scale, the supply-demand relations between all the firms in the economy do not change. We can write the grown economy completely in terms of the smaller economy if we substitute $n + \delta$ for n and $n\beta/(n + \delta)$ for β. Thus, from the same proof as in the previous section, we can assert that the symmetric equilibrium is, again, stable.

So we see that when new firms come into existence and settle in region S, the least modern firms in that region can migrate to N to restore the symmetric equilibrium. If this happens, a pattern emerges: there exists a region where new, modern production takes place, and a region where older technologies are used in production. As the economy grows, older firms are relegated to the region where the other old firms are, while new firms settle close to other new firms.

Note that we have not assumed anything about the relative productivity of either type of firm, and that we cannot arrive at any conclusion about the relative well-being of either region. Once the symmetric equilibrium is restored, that is exactly what it is: each region produces and consumes equally much of different, but equally valuable goods.

However, in practice there is often a difference between new and old industry. To be on the growing end of an economy carries some advantages, like a more dynamic labor market, positive external effects on R&D, and the presence of an environment of opportunities. In that sense, this model has a bleak message for the region harboring 'old' firms: even with continuous new entry, it will not escape its predicament through the forces of economic equilibrium. Instead, these forces seem to conspire to keep the difference between the two regions as large as possible.

A point to note is the implicit assumption that we make about the costs of moving a firm to another region. In this model, even though simple trade in final and intermediate products is subject to transaction costs, we assume that moving a firm can be done costlessly. This may be a fair approximation, if we take into account that the decision to move the firm to another region will affect the economic conditions of the enterprise for a long time, making the costs of moving relatively small. Alternatively, we could assume that firms do not actually move, but are subject to bankruptcy when their region becomes overcrowded. Their patent then becomes available and a similar firm re-emerges in the other region, where costs are lower.

If we would make the opposite assumption (large costs of moving) and instead imagine firms sticking to their region no matter what, a process of leapfrogging could emerge in this model. In this process, new firms come to the S-region until wages there have risen to a point where it is more profitable to locate in the N-region and import all intermediate goods. As soon as the first new firm has done this, region N becomes the new center of growth, leading to rising relative wages until the process again reverses. This scenario is harder to model because the different equilibria would be asymmetric, making the analytical solution of the model more involved than it was above. The characteristics of the solution would be very different: each region would see a cyclical movement in its wages and its relative attractiveness to new firms.

3.4 Conclusion

In this chapter, we elaborate on the economic geography model in which ties between firms consist of input-output relationships. When there are positive transport costs, firms will locate close to each other to minimize the costs of their intermediate inputs. A countervailing force is the labor market, where the wage rate rises in regions where many firms cluster, raising again the average costs of inputs.

When we divide firms into sectors, the agglomerating and dispersing forces can be separated. In section 3.2, we show that two groups of firms that do not use each others' products intentionally move away from each other. That is, firms that have no reason to share a region will not do so, because of the labor market pressures of other firms. Instead, they will cluster with firms from their own sector. This behavior can be seen in practice when multinational corporations decide to move manufacturing to low-wage countries. Manufacturing, in these companies, does not need the services offered in industrialized Europe or the US. Raw materials for manufacturing, on the other hand, can be obtained cheaply in low-wage locations.

We look at static equilibria in the first two sections and turn to economic growth in section 3.3. There we assume that technical progress exogenously drives growth, and that ties between firms are stronger when they are created around the same period. It turns out that, for our assumed IO-function, new firms like to enter into the region where other recent arrivals have also located. As time goes by and the economy grows further, other firms in that area are relegated to the a region with older industries. As such, there exists an innovative and a lagging region. Both are equally productive and enjoy equal welfare.

3.A Firm demand and supply

We show the validity of (3.12) by solving it for the price level p. It will turn out that this is the price level that we actually used (which is in formula 3.6).

First we write production as

$$\begin{aligned} z_i &= \left(\frac{L}{n}\right)^{\alpha} Q^{1-\alpha} \\ &= \left(\frac{L}{n}\right)^{\alpha} \left(\frac{L}{n}\frac{1-\alpha}{\alpha}\frac{1}{p_Q}\right)^{1-\alpha} \\ &= \frac{L}{n}\left(\frac{1-\alpha}{\alpha}\right)^{1-\alpha} p^{\alpha-1} n^{\frac{\alpha-1}{1-\sigma}} \qquad (3.22) \end{aligned}$$

where the second step is from (3.7).

Intermediate demand is

$$Qn^{\frac{1}{1-\sigma}} = \frac{L}{n}\frac{1-\alpha}{\alpha}\frac{1}{p_Q}n^{\frac{1}{1-\sigma}}$$

and final demand

$$\begin{aligned} L\frac{y}{p_Q}n^{\frac{\sigma}{1-\sigma}} &= (L+n\pi)\frac{1}{p_Q}n^{\frac{\sigma}{1-\sigma}} \\ &= \left(L+\frac{npz_i}{\sigma}\right)\frac{1}{p_Q}n^{\frac{\sigma}{1-\sigma}} \end{aligned}$$

Adding these two and simplifying, we get total demand

$$\mathrm{TD} = \frac{L}{n}\frac{1}{p\alpha} + \frac{z_i}{\sigma}$$

Into this expression, put (3.12) by requiring that $\mathrm{TD} = z$.

$$\frac{L}{n}\frac{1}{p\alpha} = \frac{\sigma-1}{\sigma}z.$$

We then use the expression from (3.22) for z and simplify:

$$\begin{aligned} \frac{L}{n}\frac{1}{p\alpha} &= \frac{\sigma-1}{\sigma}\frac{L}{n}\left(\frac{1-\alpha}{\alpha}\right)^{1-\alpha} p^{\alpha-1}n^{\frac{\alpha-1}{1-\sigma}} \Longrightarrow \\ p &= \left(\frac{\sigma}{\sigma-1}\right)^{\frac{1}{\alpha}}(1-\alpha)^{\frac{\alpha-1}{\alpha}}\,\alpha^{-1}n^{\frac{1-\alpha}{\alpha(1-\sigma)}} \end{aligned}$$

Q.E.D.

Chapter 4

Input-output relations in economic geography

4.1 Introduction

An unhindered flow of trade between countries is beneficial to all parties. The most fundamental result of Ricardian trade theory depends on the differing proficiency of different nations in the production of various goods. In addition to the Ricardian gains from trade, Samuelson (2001) reminds us that there exists a *Sraffian Bonus* (Sraffa 1960) when imported goods can be used as intermediate inputs. In a model without costs of transport and with a simple linear production technology, free trade leads to less expensive final products for consumers, but also to cheaper inputs for firms. The latter causes world industry to become more productive when trade is allowed, leaving everybody better off. Like the standard Ricardian gains from trade, this result depends on comparative advantage and holds even when one country dominates the other in absolute productivity in all sectors.

In spite of these theoretical recommendations, not everybody favors free trade. Small, peripheral or underdeveloped countries often choose to close their borders to protect domestic industry. They fear that competition from abroad will be too strong, because foreign competitors are either larger, more established or both. Indeed it does not take much by way of changing the assumptions to throw a spanner in the works of trade theory and diminish the appealing result above. It can be argued that Venables (1996a) does just that. His model has nonzero costs of transport and a nonlinear production technology. There are two sectors, industry and agriculture. Firms in the industrial sector use each other's products as intermediate inputs and because of international price differences in these inputs, a possible outcome of the model is the agglomeration of all industry in one country.

In this chapter, we look at a models that is similar to that of Venables

(1996a). There is an important distinction, however: we allow for more than one industrial sector, and the demand for intermediate inputs can differ between sectors. This more general model will allow us to make observations on the likelihood of complete agglomeration of industrial activity.

4.1.1 Related theory

The model by Venables cited above belongs to the family of Economic Geography models. The central tenet of this body of theory is that agglomeration of activity can be an economic equilibrium, sustained by complementarities that exist in the production process. In the first chapter, we discussed different mechanisms for these complementarities, which formalize the concept of forward and backward *linkages* between firms and the local labor market, or among firms in the same region (These linkages are similar to those discussed by Hirschman, 1958). We briefly repeat the three mechanisms:

- The home market effect, which occurs because firms demand local labor and the local labor force demands the firms' products.

  When both are mobile the interaction of their demands may result in a large local market for both labor and endproducts. In this process, workers choose to move to the agglomerated region which, in turn, becomes an attractive place of business for firms. Models of this type typically rely on a mobile workforce which is tied to its sector, *i.e.*, there are a fixed number of farmers and manufacturing workers. An early example is found in Krugman (1991a).

- Linkages through intermediate goods. When the assumptions on worker mobility are reversed, we obtain a model that applies in an international context. Workers are not allowed to move to a different region, but can choose in which sector they work. Agglomeration can occur in the sense that one region gets all of the industrial activity while the other must import its industrial products. Firms choose to be in the same region because they depend on each other for intermediate goods, the availability of which may offset higher wage costs. This model was developed by Venables (1996a). However, the notion that intermediate goods play a role in the agglomeration process can be traced back all the way to Marshall (1920).

- Finally, we identified a class of models where the agglomeration mechanism is the result of a third sector, where R&D is conducted. This type of model was discussed by Martin and Ottaviano (1996b).

Each of these three classes of models uses the principle of complementarity to explain agglomeration, but specifies a different channel through which it operates. It depends on the situation that is studied which one of these channels, and thus which one of these models, is most relevant.

In this chapter, we will take a closer look at the second type of economic geography model. Intermediate products are the channel for complementarities here: firms benefit from each other's presence as each producer uses the products of other firms as an intermediate input. When transport costs are nonzero, agglomerations can occur because the advantage of being close to intermediate supply and demand outweighs possible disadvantages, such as a higher wage level. We discuss our choice for this model below.

4.1.2 Intermediate goods, trade, and ties between industries

There are several reasons why the intermediate goods-type of model[1] forms an interesting subject for further research. Its assumptions specify that workers cannot relocate to another region, but *can* change the sector in which they are active. These assumptions are appropriate when we want to study trade between different nations. As Fujita et al. (1999, p. 239-240) show, mobility between different nations may not be zero, but it is close to zero. In the European Union, legislation has been passed to facilitate the movement of workers between the different member states. However, the number of workers who actually migrate to another country is very low. In the year 2001, for instance, only two percent of EU nationals worked in another member country (Migration News 2001). Language problems and "soft barriers," such as differences in pension systems and tax codes, cause the immobility. Meanwhile, it is clear that some European regions are more agglomerated than others: the concentration of activity in the "hot banana" that lies between London, the Ruhrgebiet and Northern Italy[2] is much higher than that in some of the more peripheral European countries. These facts clearly call for a model that features both immobility and agglomeration.

If we want to gather more evidence about the relevance of the Venables model for international trade, we can look at data on flows of trade between industrialized and non-industrialized nations and decide whether its characteristics match the model's predictions. Given that in the model, each firms uses *every* industrial product as an input, it predicts that the flows of trade between industrialized countries contain at least some intermediate products. Final products should flow from industrialized countries to both the periphery and to other industrialized countries. Finally, the periphery

---

[1] From now on, we will use the term 'Venables-model' in this text, recognizing the author of the first economic geography-model which used intermediate goods as a channel for complementarities (Venables 1996a).

[2] This area is also known as the "blue banana," after its discovery on a colored map of Europe (ESPON 2003, p. 69).

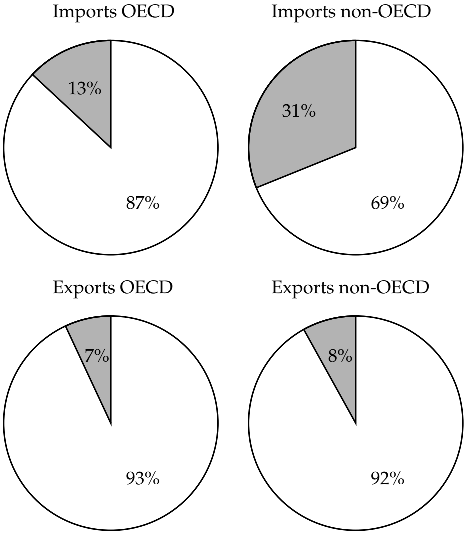

Figure 4.1: Trade data for Germany in 1999, from and to OECD- and non-OECD countries, from OECD (2000). The smaller, grey slices are the share of 'basic' products (sections 0–4 in the standard international trade classification), while the remaining, white slices are 'industrial' products (sections 5–8). Sections 0 through 4 are, respectively: food and live animals, beverages and tobacco, crude materials (inedible), mineral fuels, lubricants, animal and vegetable oils, fats and waxes. Sections 5 through 8 are chemicals and related products, manufactured goods, machinery and transport equipment.

pays for its imports with non-industrial goods, which should make up the flow of trade from them to the industrialized countries.

To find out about the relevance of the model we could inquire about the accuracy of its predictions. Ideally, we would gather data about trade between a number of regions, agglomerated and peripheral, and find out about the share of final, intermediate and basic, non-industrial products. Unfortunately, data is not gathered using these definitions; we will have to make do with a first-cut approximation, in which we group the more-agglomerated regions and compare them to their less-agglomerated cousins, dividing up the flows of trade in more- and less-basic goods.

An example of such data is in figure 4.1. It shows the characteristics of the imports and exports of Germany in the year 1999. The flows of trade are split into four categories: firstly, we separate out trade with OECD nations, which we take as a rough approximation of trade with industrialized countries, and contrast it to the trade with non-OECD nations. Secondly, we use a crude categorization of 'basic' and 'industrial' goods, corresponding to different sections of the trade statistics. Our 'industrial' goods will be a proxy for the final and intermediate goods of the model. The 'basic' goods can be produced without industry and correspond to the 'agricultural' sector of the model.

From the figure, we see that our predictions come through in a relative sense: a relatively large share of the goods that Germany imports from countries outside the OECD is basic, compared to the imports from fellow-OECD countries. Most exports are of an industrial nature.[3] But while the data are roughly consistent with the model, there are some differences. We would have predicted imports from non-OECD countries to be all basic or agricultural goods and the three other streams to be largely industrial. In fact, industrial products are a non-negligable part of imports from non-OECD countries. We can identify several reasons for this inconsistency.

First of all, if we assume that the model is true, we might explain the differences between its predictions and the data by measurement error. Our division between OECD- and non-OECD countries does not coincide exactly with agglomerated and non-agglomerated areas. Also, the different sections of the trade statistics that we have used do not exactly match *industrial* and *basic* products in the model's sense. Moreover, some goods such as oil double as both industrial and basic. Finally, because we have measured trade in dollar terms, we can expect the more expensive industrial goods to carry more weight than they would have in terms of weight or volume.

However, at the root of these inconsistency could also be a problem with the model's simplicity. Maybe some industrial firms *did* establish in the non-agglomerated region, where wages are lower. These firms could depend less than average on intermediate products. Because model does not allow for different kinds of industrial firms, this nuance is not a part of its predictions. It is part of this shortcoming that we will try to remedy in the current chapter.

When we look closer at the streams of trade between industrialized countries, we see another shortcoming of the basic Venables model. For instance, look at the data in figure 4.2. It is a breakdown of section 7, or *machinery and transport equipment*-trade between Germany and the Netherlands. Each way, approximately 11 billion US$ worth of goods is shipped. However, the intra-section division is completely different: while Germany exports mostly type 7.8, or *road vehicles*, they import mostly 7.5, or *office machines and computers*.[4]

---

[3] It could of course be argued that the large share of industrial goods in the imports from non-OECD countries is an indication that our approximation of Industrial and non-Industrial countries is wrong. While a dedicated researcher could probably find better data, the very point of this exercise is to show that simple notions of countries as being either completely industrialized or devoid of any industry are wrong.

[4] Other categories: 7.1: Power generating machinery and equipment, 7.2: Machinery

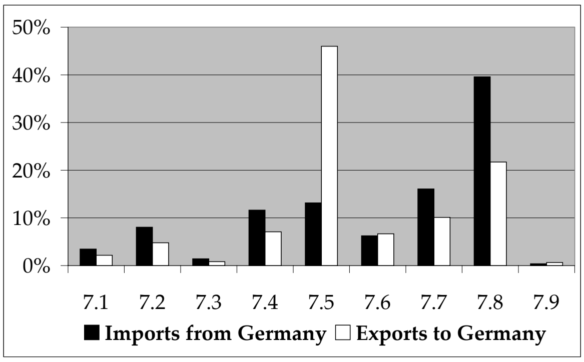

Figure 4.2: Section 7 trade between the Netherlands and Germany, from OECD (2000). Section 7 is *machinery and transport equipment*. The Netherlands exports mainly 7.5-type goods, *office machines and automatic data processing machines*, to Germany while German exports to the Netherlands contain a large share of section 7.8, *road vehicles*.

This is where the assumption that industry is a homogenous sector again becomes impractical. We know that there exist different sectors, even though in the model they have been lumped together for convenience. Now they turn out to be concentrated in different regions: it appears that firms in the automotive industry are more concentrated in Germany, while the makers of office appliances have concentrated in the Netherlands. This suggests that there is an agglomerating force *within* industries, as well as the force *between* industries that our model predicted.

The existence of such a force is discussed by Krugman (1991a), who gives three reasons for its existence (this is the Marshallian trinity, as they were originally proposed by Marshall 1920). The first two reasons argue that concentration is the effect of specialized labor markets and of externalities. The third explanation, is more in line with the Venables model: the reason for agglomeration within a sector is that inputs *specific to an industry* are available in greater variety and at a lower cost when firms are close together. This is at odds with the assumption that industries are homogenous, since it implies that they use different sets of inputs. Hence, it may be efficient for these industries to agglomerate into different regions.

Indeed, in real life some firms depend on a large amount of intermediate goods and some are less dependent on intermediates, and the types of intermediate goods are known to vary. In the current model, there are only two types of firms, one of which (agriculture) uses no intermediate goods

---

specialized for particular industries, 7.3: Metalworking machinery, 7.4: General industrial machinery, 7.6: Telecommunications and sound equipment, 7.7: Electrical machinery, 7.9: Other transport equipment.

while the other uses a bundle comprising all products.

4.1.3 Plan for the chapter

In this chapter, as in chapter 3, we will develop a model of the Venables type. In Section 4.2 below, we will briefly review the Venables (1996a) model and add an extension which allows us to model different industrial sectors. In this chapter, the sectors will be discrete so that it is possible to construct a traditional input-output table that specifies the flows of trade between sectors. We will find that there are several possible equilibria in the extended model, apart from agglomeration and dispersion. In contrast to Krugman and Venables (1996), who assume symmetry between sectors and Fujita et al. (1999), who only discus a small number of special cases, we will to study the effects of different IO-structures in their entirety, concentrating on the boundaries between different types of equilibria.

Section 4.3 discusses the possible types of solutions. We will see that 'catastrophic' changes in the agglomeration of firms can happen when IO-parameters shift marginally. This effect has previously been shown for the transport costs parameter.

Finally, section 4.4 concludes and looks ahead to possible applications of the results.

4.2 The Venables model, and an extension with discrete sectors

4.2.1 One industrial sector

In this section we discuss the Venables (1996a) model with decreasing returns in agriculture. The model is also briefly discussed in chapter 14 of Fujita et al. (1999). A few aspects have been changed to facilitate later extensions, these changes will be noted in the text. In the initial model, we will assume that the world consists of two regions and there are two sectors of production, an agricultural sector and an industrial sector. Later on, we will generalize both the number of regions and the number of sectors.

Both regions are similar in principle, even though their state in equilibrium may be different. Each region has a fixed, immobile labor supply, whose members may choose to work in the agricultural sector or the industrial sector. We describe the production facilities, which are the same in both regions. Total production in the agricultural sector depends only on L_A, the amount of labor used, and is equal to L_A^β $(0 < \beta < 1)$. Total agricultural production is divided among all workers in the sector, so that their income equals

$$w = L_A^{\beta-1}. \tag{4.1}$$

In this setup, the wage is greater than the marginal product of labor. This follows because of an implicit assumption that the other production factor, land, is freely available. This way, we do not have to introduce a separate class of landowners. Note that $\beta - 1$ is a negative number, so that the wage rises as the number of farmers drops. Labor is in fixed supply, but can move freely from one sector to another. Therefore, wages are equalized between sectors and the expression in (4.1) is the wage for the entire region. This means that a region with a large industrial sector and small agricultural sector will have relatively high wages.

In the industrial sector, production of a firm i is a function of the applied amounts of labor and intermediate products. Each firm makes a unique product, but all firms share the same production technology; we will therefore omit the subscript i below. The production function for a firm is

$$y = \frac{1}{\phi \cdot \theta_\alpha} L^\alpha Q^{1-\alpha} - F. \tag{4.2}$$

In this function, L is the amount of labor applied and Q is an aggregate of intermediate products. Labor's share α lies between 0 and 1. There are fixed costs F of production which are incurred in the final product. Finally, the constant scaling factor consists of two terms, a positive scalar ϕ and the positive number θ_α, which is equal to $(1-\alpha)^{\alpha-1}\alpha^{-\alpha}$.

From the above production function, it follows that marginal costs are equal to

$$MC = \phi w^\alpha G^{1-\alpha} \tag{4.3}$$

We will normally set $\phi = 1/\theta_\alpha$, so that function 4.2 is of a simple form.

There is a continuum of firms that employ the above production function. As mentioned, their products are unique and we assume limited substitutability between them. All products serve as both final and intermediate. In order to aggregate these products into a single intermediate input Q, we use a method first employed by Ethier (1982): the size of Q is a Dixit-Stiglitz (1977) aggregate of all the different varieties. This amounts to saying that

$$Q = \frac{1}{\psi}\left[nq^{\frac{\sigma-1}{\sigma}} + n^*(q^*)^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}}. \tag{4.4}$$

In this equation we introduce our convention to separate the two regions by designating one as *home* and one as *foreign*. We add asterisks to the variables of the *foreign* region. The number of firms, and thus the number of distinct products, in *home* is equal to n, and there are n^* firms in *foreign*. When we bring together a quantity q of each of the n home products and a quantity q^* of each foreign product, formula (4.4) shows that a bundle of intermediate goods of size Q results. This variable plays two roles: firstly, it serves as a measure of intermediate product as in formula (4.2). Secondly, we assume that consumers demand industrial products in bundles of Q. This means that both firms and consumers have an elasticity of substitution of σ between the different industrial products. As for consumers, we furthermore assume that they spend a fixed fraction μ of their income on industrial products and a fraction $1-\mu$ on agricultural products. The positive scalar ψ, finally, determines the level of costs of the intermediate bundle. Making it larger means that intermediate goods become more expensive relative to the other factor of production, labor.

We must specify if products of one region are available in the other. We take a dual approach in this matter: agricultural products can move freely at zero cost, but industrial products are subject to iceberg trade costs: only a fraction τ (where $0 < \tau < 1$) of shipped goods arrives—alternatively, the price of goods from another region is $1/\tau$ times the f.o.b. price.

As we saw in section 2.2.2 in the previous chapter, the aggregation in formula (4.4) implies a price index G for the bundle Q, equal to

$$G = \psi \left[np^{1-\sigma} + n^*(\frac{p^*}{\tau})^{1-\sigma} \right]^{\frac{1}{1-\sigma}} . \tag{4.5}$$

As expected, there holds that an amount E will buy E/G bundles of Q. In formula 4.5, p and p^* are the prices of home and foreign products in their own region. Notice that because of transport costs, the price for *foreign* goods in *home* is higher than p^* by a factor $1/\tau$.

Formula (4.5) implies that firms in the same region set the same price. Knowing that they are in monopolistic competition, all firms employ markup pricing in equilibrium, setting their price at

$$p = \frac{\sigma}{\sigma - 1} w^{\alpha} G^{1-\alpha}. \tag{4.6}$$

This price is a markup over marginal costs. A similar relationship holds for the foreign region. Notice from (4.5) and (4.6) that there is a circularity between price p and price index G, which cannot be solved analytically. When parameter values are known, however, both G and p may be computed by numerical means.

Clearly, there is a dependence between the two regions, which works as follows. As already stated, labor is tied to its own region. We stipulate that the same holds for the ownership of firms, so that the only interaction between regions takes the form of trade.

As in chapter 1, we have to assume something about the relative speed of adjustment of workers and firms. Workers move between the two sectors in response to wage differences, firms enter or leave the market in response to profits or losses. We will retain our assumption that workers move faster than firms, so that at all times the wage is equal in both sectors, and given by equation (4.1). For firms, the zero-profit condition is $y = (\sigma - 1)F$. We will study the model out-of-equilibrium, so that this condition does not hold everywhere in this chapter.

Optimization by firms implies that labor and intermediate products will be employed such that their marginal costs equal marginal benefits, so that the quantity Q/L can be computed from the relative price w/G. The relative use of intermediate product can be found both in- and out-of the zero-profit equilibrium. When profits are zero, the applied amount of labor is a function of parameters and relative prices:

$$L = \sigma F \alpha \left(\frac{w}{G}\right)^{\alpha - 1}. \tag{4.7}$$

Now that we have specified the model, we can think about possible states of equilibrium. We know that each region must have an agricultural sector, because productivity in this sector goes to infinity when its size goes to zero (formula 4.1). At least one region has an industrial sector, as a positive fraction of income is spent on industrial goods no matter what. If both regions have industrial firms, there will be intra-industry trade. As a result of the Armington assumption both firms and consumers will demand domestic and foreign products, though less when trade costs are large.

Because both regions are exactly similar in all aspects, it would appear that a situation where both have the same number of firms might be an equilibrium. Simulation confirms that this, sometimes, is the case. However, the model allows for other solutions as well: for certain sets of parameters, all industry can agglomerate into a single region, leaving the other with only an agricultural sector. It is this result, first obtained in the work of Venables (1996a), that one should be mindful of before opening one's region to trade: it is quite possible that all domestic industry will be lost to the neighbor's industrial core. Thirdly, for a small subspace of parameters, an asymmetrical equilibrium is possible in which both regions have a positive, but different number of industrial firms.

We will use numerical results to show that an agglomerated equilibrium can be stable. It is quite easy, *given that a single industrial core exists*, to compute what it must be like. With $n^* = 0$, expressions such as formula (4.5) become less tangled. We can then compute a solution for all variables.

Using this solution, we evaluate the position of a hypothetical single (small) firm in the foreign region. We used the same method of evaluating the agglomerated equilibrium in section 2.C above, where there was only one sector of industrial firms. This firm would be the only one of its sort, a pioneer breakaway from the core. We can compute the demand that this firm would receive, both from home and abroad, its costs and its profit. The firm would have several advantages over the firms in the agglomerated region: wages are lower and local demand is stronger. Its disadvantage is the high cost of intermediates, which all have to be shipped from the other region. Also, demand from other firms will be lower than it would have been in the agglomerated region. In the balance, these factors determine the sign of the hypothetical firm's profit. If this potential profit is negative, we conclude that our assumption of zero industry in the foreign region was correct; it would not be profitable to start any. However, if potential profits are positive we must conclude that the agglomeration of industry in the home region is unstable.

Figure 4.3 shows the potential profits of a breakaway firm as a function of transport costs τ.[5] We see that an agglomeration of all firms in one region is stable for intermediate values of τ. This is a well-known result, known as *inverted U dependence*. For very high transport costs, local demand alone is enough to sustain industries in both regions. For very low transport costs, the advantages of agglomeration are outweighed by the disadvantages of using only one workforce. In between, agglomeration is stable.

But how about the other way around? Does the stability of the agglomerated equilibrium automatically preclude a stable symmetric equilibrium? We investigate the symmetric equilibrium as follows. Using numerical methods, it is possible to compute what the symmetric equilibrium would look like, given that it exists. Both regions have an industrial sector, and all firms turn a profit of exactly zero. The regions are completely symmetric: wages, sectoral structure, import and export are the same on either side.

Using the same numerical methods, we can compute the effects of infinitely small changes in the number of firms, allowing wages and prices to adapt but momentarily suspending our assumption of zero profits. This is in line with our earlier assumption that people move faster between sectors than firms move between regions, an assumption that was employed in section 2.C. Thus, we look at the effects of small perturbations in the number of firms: what would happen if an extra firm opened up here, another firm closed there? To capture all possible perturbations, we compute the matrix of derivatives $\partial\Pi/\partial N'$, which is constructed as follows:

$$\partial\Pi/\partial N^T = \begin{pmatrix} \frac{\partial\pi}{\partial n} & \frac{\partial\pi}{\partial n^*} \\ \frac{\partial\pi^*}{\partial n} & \frac{\partial\pi^*}{\partial n^*} \end{pmatrix} \tag{4.8}$$

In this matrix, n is the number of firms in *home* and n^* the number of firms in *foreign*. The vector N stacks both, N^T being its transpose. The vector Π stacks the two profits: each firm in *home* turns an profit of π, the profit of a foreign firm is π^*. In equilibrium, both are zero by definition. Their derivatives with respect to n and n^* can be nonzero, though. From the matrix $\partial\Pi/\partial N$, which contains these derivatives, we can judge the stability of the symmetric equilibrium. Suppose a small increase in n leads to a higher profit for *home* firms, π, and that this was the *only* effect of the increase. Then such a change in n would be self-enforcing: the positive

---

[5] Note that transport costs are high when τ is low and small when τ is close to unity. The profits in figure 4.3 are the results of a numerical simulation. The other parameters in example are: $\sigma = 4$, $\beta = 0.8$, $\alpha = .37$, $\mu = 0.8$, $\psi = 1$ and $F = 0.001$.

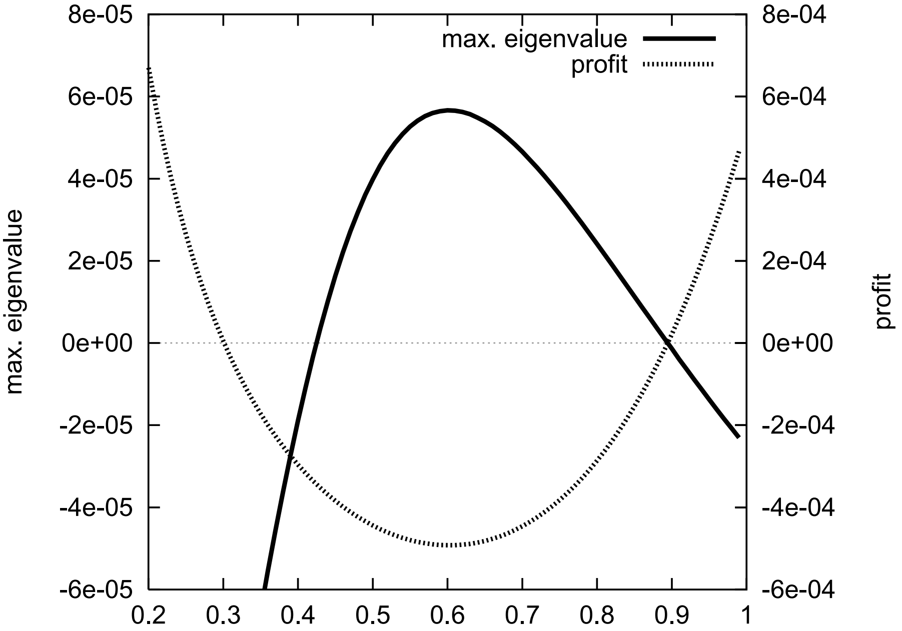

Figure 4.3: This figure shows two lines that indicate the stability of two equilibria as a function of transport costs τ (τ is on the horizontal axis). The broken line pertains to the agglomerated equilibrium. On the right-hand y-axis, it shows the potential profit of a single firm in the other region. If this is negative, the equilibrium is stable. The continuous line uses the left-hand axis and pertains to the symmetric equilibrium. It shows the maximum eigenvalue of the matrix $\partial\Pi/\partial n$, the derivative of profits to the number of firms. If there exists a positive eigenvalue, the symmetric equilibrium is unstable. Note that the lines cross the x-axis at the same point on the high-τ side, but not on the other side.

profits would draw in more firms, which would in turn cause even higher profits, *et cetera*. Similarly, a fall in the number of firms would drive profits below zero. This would encourage more firms to exit, causing profits to fall further. We can conclude that if the upper left element of the matrix is positive, and the rest are zero, the equilibrium is unstable.

By analogous reasoning, we can see that a negative upper-left element and surrounding zeros would make the equilibrium stable. However, the matrix of derivatives is rarely this simple. Instead we must generalize the above approach and examine the eigenvalues of the matrix $\partial\Pi/\partial N$. Suppose there exists an eigenvalue greater than zero. Now suppose that a perturbation to the number of firms, $\partial(n, n^*)$ occurs that just happens to be a multiple of the corresponding eigenvector. The resulting changes in profits $\partial(\pi, \pi^*)$ would have the same direction as the change in the number of firms, $\partial(n, n^*)$. This change would be reinforcing, as it was in the scalar example above. It appears that the presence of one of more positive eigenvalues signals that the symmetric equilibrium is unstable.

This assertion can be made more rigorous. There presumably exists a relationship between profits and the number of firms which we have left implicit, stating only that firms enter or exit until profits are zero. Suppose for a moment that this relationship is linear, say,

$$\dot{N} = \frac{\partial N}{\partial t} = a \cdot \Pi \tag{4.9}$$

with a a positive scalar and N and Π vectors, as above. Now consider the equilibrium numbers of firms $\tilde{N}$ and the variable $N' = N - \tilde{N}$. The equilibrium value of N' is zero.

Now we linearize our model around $\tilde{N}$. By (4.9), there holds that

$$\begin{aligned} \dot{N}' &= a \cdot \Pi \\ &\approx a \cdot \frac{\partial \Pi}{\partial N^T} \cdot N' \end{aligned}$$

This is a homogeneous linear differential equation in N', whose stability properties are determined by the matrix $a \cdot \partial\Pi/\partial N^T$. If the real parts of its eigenvalues are all negative, the solution is asymptotically stable (see, for instance, Brock and Malliaris 1989, p. 66). As the sign of the eigenvalues does not change because of the positive scalar a, the original result still holds.

How does this analysis relate to the one-dimensional computations of break- and sustain values of transport costs that was introduced in section 2.C? Recall that we made a number of assumptions in that section to calculate the break-point, the value of transport costs at which the symmetric equilibrium breaks down: first of all we started with a symmetric equilibrium, and we assumed that changes in the number of firms in the two regions were opposite and of equal size. That is, we studied a perturbation in the number of firms in one region, and simultaneously looked at the opposite perturbation in the other region. Because there was only one sector of industrial firms, the perturbation took place along a single dimension. This made the analysis easy to handle, as only one derivative had to be computed. The fact that the equilibrium was symmetric allowed us to simplify the model to the point where an analytical solution to the derivative could be found.

In the current setup, the increased complexity of the model makes it much harder to arrive at a similar simplification. For one, there may be more than one sector of industrial firms and the sizes of the different sectors do not have to be symmetric. Furthermore, if we want to rule out destabilizing changes in the number of firms, we now have to look at all possible changes in a higher-dimensional space. With two sectors, for instance, the equilibrium may be stable with respect to an increase in the number of firms in both sectors, but unstable against an increase in one, and a decrease in the other sector. This calls for a more general method, in which all possible changes in the number of firms are analyzed at once. An added benefit is that we no longer need to concentrate on symmetric changes only.

Using the eigenvalues of the matrix in (4.8), we can quickly identify destabilizing perturbations. However, this more general analysis has the drawback that we rely on a numerical approximation to the derivative, where previously an analytical solution could be found. This has to do with the increased complexity of the model, where now even a symmetrical equilibrium can be less than straightforward as the relative number of firms in the two sectors depends on the value of the input-output matrix. This precludes a sufficient simplification of the problem.

In figure 4.3, we have plotted the maximum eigenvalue for the matrix $\partial\Pi/\partial N$ (given that there is a symmetric equilibrium) for the same range of transport costs as before.[6] We see that on the high end of τ, the symmetric equilibrium is unstable for exactly the same range of values for which the asymmetric equilibrium was stable, and *vice versa*. On the low end of τ, there exists a region where both equilibria are stable. In this part, history decides which equilibrium attains. We conclude that in general, the symmetric equilibrium is stable at the high and low transport costs, but not in between.

4.2.2 Types and determinants of equilibrium

The equilibrium that attains in this model is the result of opposing forces of agglomeration and dispersion. A force of agglomeration is the use of intermediate products that are subject to transport costs: when dependent on these products, it pays to be close to their suppliers. A force of dispersion are the scarce laborers: when many firms pack into one region, wages will go up and settling elsewhere becomes more attractive. We discuss the factors that affect the balance between these forces and look at the equilibria that result.

Firstly, to illustrate the dispersive effect of elastic wages, we examine the counterexample $\beta = 1$. In this configuration there are constant returns in agriculture and wages are equal to unity in both regions regardless of the whereabouts of the industrial sector. The results of a simulation with this model are in figure 4.4, whose setup is similar to figure 4.3 on page 90.[7].

---

[6] As before, these values are the result of a numerical simulation, in which the Hessian matrix 4.8 is approximated.

[7] In the simulation of figure 4.4, we have set the share of consumer income spent on manufactures (μ) to 0.4. This way, the industrial sector is small enough to agglomerate into

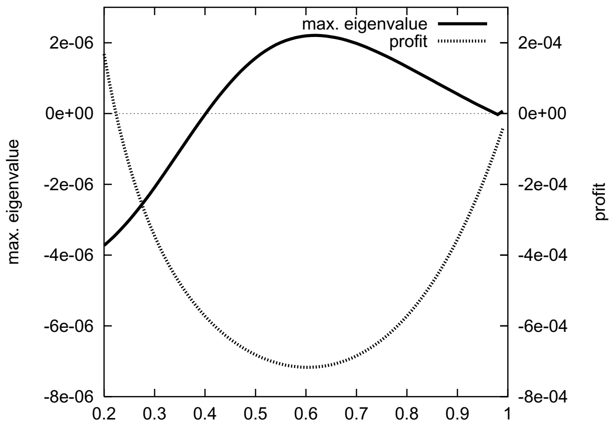

Figure 4.4: The stability of two types of equilibrium as a function of transport costs τ (horizontal axis). See figure 4.3 for details. In this simulation, $\beta = 1$ so that wages are constant. The share spent on manufactures, μ, is set equal to 0.4.

We see that the maximum eigenvalue of the matrix $\partial\Pi/\partial N$ is larger than zero for almost all values of τ, indicating that the symmetric equilibrium is unstable. At the same time, the profits of a firm that would defy the agglomerated equilibrium are negative for a large range of τ, indicating that total agglomeration is a stable outcome. These results confirm that elastic wages form a force of dispersion: without them, agglomeration is almost inevitable, as there are no disadvantages to clustering into one region.

After we reinstate decreasing returns in agriculture, it stands to reason that the relative importance of the factors labor and intermediate products will be an important determinant of the type of equilibrium that is found. In our model, a measure of this relative importance is the parameter ψ. This can be seen from formula 4.5: when ψ is high, productivity in the making of intermediate goods is lower, rendering them more expensive. This causes producers to shift to labor as an input factor and thus increases the effect of wage differences.

Using equations 4.5 and 4.3, we can compute how ψ factors into marginal costs. Combining the two expressions shows that levels marginal costs are proportional to $\psi^{1-\alpha}$. This factor serves to magnify the effect of w. A high value of ψ means that agglomeration, and its accompanying wage difference, become less likely. A low value of ψ means that wage differences become less important and can be overcome in favor of agglomerative forces.

A number of simulations where different values of ψ are used can be

---

one region. The original, larger value of μ would have caused expulsion of agriculture from one of the two regions, presumably driving up the wage after all. For clarity, we avoid this complication.

seen in figures 4.5 through 4.7. In the first figure, we have set ψ to 0.1, making intermediate goods cheaper relative to labor. The change shows up mainly in the level of the maximum eigenvalue of $\partial\Pi/\partial N$, which is much lower than before. Interestingly, the sign of the maximum eigenvalue as a function of τ hardly changes, leaving the relation between transport costs and equilibrium almost the same as in figure 4.3. It appears than although the incentive to move away from a symmetric equilibrium is smaller, it is still *positive*.

Things are entirely different in figures 4.6 and 4.7, where $\psi = 2.8$ and $\psi = 20$, respectively. Making intermediate products much more expensive enhances the dispersive power of wages to the point that only for a very small portion of τ-space, agglomeration is stable and dispersion is unstable.

So far, we have talked only about complete agglomeration and complete symmetry, although we mentioned a third possible equilibrium. That type of equilibrium occurs for small subset of all possible combinations of ψ and τ. Observe that in figure 4.6, where $\psi = 2.8$, for τ just below 0.8 both lines lie *above* the x-axis, indicating that both the symmetric and agglomerated equilibrium are unstable. Figure 4.8 shows a close-up of that part of the τ-axis. In this case, we find that the only stable equilibrium is one where both regions have some industry, although one region has a smaller number of firms than the other.

The third equilibrium, which we will call the 'overflow' equilibrium[8] plays a role when we look at situation where transport costs steadily decrease. A world where τ becomes larger, ultimately reaching unity, was first discussed by Krugman and Venables (1995). They showed that in such a scenario, the equilibrium will jump from one state to another, as stability changes. The possibility of an overflow equilibrium as an intermediate stage between agglomeration and symmetry precludes such jumps. The overflow equilibrium only occurs for certain values of ψ, though.

As a theoretical result, the inverted-U dependence on transport costs is both surprising and useful. It shows that the economic geography-type of models can be applied in the context of international trade. It also allows the broad insights that are given in Krugman and Venables (1995), among others. However, as a tool for empirical analysis, the model is too coarse. It assumes two sectors, agriculture and industry, where the latter is completely homogeneous. This assumption does not do justice to the complicated relations that often exist between different firms in the 'developed part' of an economy. To understand the complexities of relations between different countries, we must be able to characterize industries as upstream or downstream, for instance. A natural extension to the model would therefore be to specify the input-output relations that exist between

---

[8] In the overflow equilibrium, the agglomerated region lets some of its firms flow into the agricultural region, but remains the dominant seat of the industrial sector.

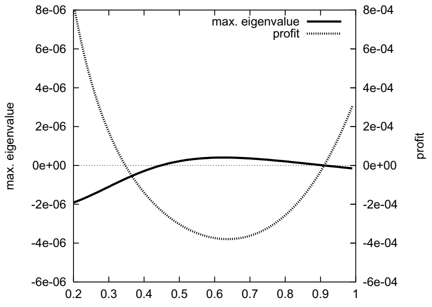

Figure 4.5: Stability of two types of equilibrium as a function of τ, with $\psi = 0.1$. This makes labor much more expensive than intermediate products.

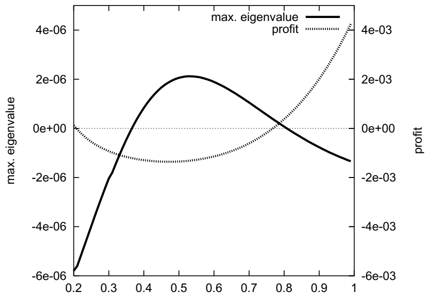

Figure 4.6: $\psi = 2.8$. Intermediate products are more expensive than labor. Note that there is an area where both equilibria are unstable, close to $\tau = 0.8$. See figure 4.8 below.

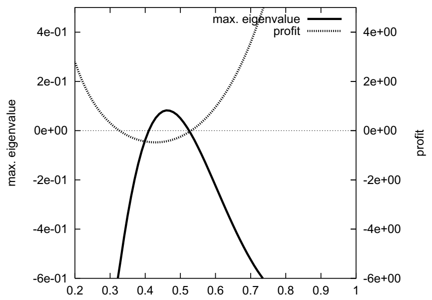

Figure 4.7: $\psi = 20$. Intermediate products are much more expensive than labor.

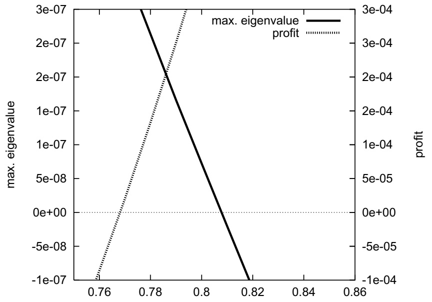

Figure 4.8: $\psi = 2.8$. A closeup look at figure 4.6, which reveals that there exist values of τ where both equilibria are unstable.

different industries in an input-output (or IO) matrix.

We shall use the term *IO matrix* to refer to the set of parameters that indicate how intermediate products from different sectors enter the production function of the various firms. It has a natural empirical counterpart in the *IO table* of an economy. This table, which is regularly constructed for all major economies, specifies the volume of trade between the different sectors. As such, it is an indication of the strength of linkages between those sectors, given that these linkages work through the trade in intermediate goods.

The extension of the model in Venables (1996a) with an IO matrix is taken up in a number of papers, including Krugman and Venables (1996) and Venables (2000). A useful summary of the results is given in chapters 15 and 16 of Fujita et al. (1999). Their main results are two. Firstly, if you assume a form of labor-augmenting technological growth, plus a number of sectors connected by a fairly general IO matrix, an interesting growth process follows. In the beginning, all the industry is agglomerated in one region. Once this region is too small to hold all industry, some sectors make the jump and agglomerate in the second region. This pattern continues, and it suggests a mechanism in which the growth process is punctuated by sudden changes in the economic structure. Differences in the IO matrices are kept to a minimum—the analysis shows that upstream sectors are the first to leave a region, as are those with the weakest links to other industries.

A second main result is that in a model with two regions, no growth and only industrial production, there are two possible equilibria, depending on the costs of transportation. The two sectors can either both choose to settle in both regions, leading to a mixed equilibrium, or the regions can become specialized, each being the host to only one sector. This model is used to explain the fact that many industries in the US are concentrated, while the same industries appear in many countries in the EU. The authors show that this phenomenon can be traced back to lower costs of transportation in the New World.

4.2.3 A model with discrete sectors

With this model, we will try to get some insight into the different types of equilibrium that obtain when we vary the IO matrix. In order to limit the possible number of equilibria and keep the analysis manageable, we only look at the simplest possible setup: a situation where there are two industrial sectors and two regions. This will allow us to present the results in a graphical manner later on. An extension to more sectors is straightforward and pursued in chapter 15 of Fujita et al. (1999), among others. We have made a different generalization in chapter 3, where we discussed a continuous IO structure.

We assume two regions and two kinds of firms, agricultural and industrial. The industrial firms are divided into two sectors. The products of all the different firms are consumed in both regions by agents who maximize utility,

$$U = A^{1-\mu_1-\mu_2} Q_1^{\mu_1} Q_2^{\mu_2} \tag{4.10}$$

where A is consumption of the agricultural good and Q_i is the consumption of products of (industrial) sector i ($i = 1, 2$). As before, we assume that agricultural good is homogeneous and freely tradeable across regions. It will serve as the numéraire. As follows from (4.10), the agricultural sector receives a fixed fraction $1 - \mu_1 - \mu_2$ of each agent's income.

The industrial goods are heterogenous again, and each sector is subject to monopolistic competition. The aggregation goes according to

$$Q_i = \left(n q_i^{\frac{\sigma-1}{\sigma}} + n^* (q_i^*)^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} \tag{4.11}$$

for $i = 1, 2$. This equation is similar to (4.4), but it now operates on the sectoral level. Notice that we assume that the values of σ are equal for the two sectors. This is not essential to the results, but does simplify the analysis considerably. Notice also that, compared to (4.4), we have left out the scaling parameter ψ. This parameter, which is meant to vary the level of costs of intermediate goods, will be reintroduced at the appropriate level below.

The different Q_i's are themselves aggregated by consumers (as in formula 4.10 above) and serve as bundles of intermediate products. The aggregation of Q_i's into a factor of production is sector-specific and follows from each sector's production function. This function is

$$y_i = \frac{1}{\phi \cdot \theta_\alpha} L_i^\alpha \left[\frac{1}{\psi_i \cdot \theta_{\eta,i}} Q_{i,1}^{\eta_i} Q_{i,2}^{1-\eta_i} \right]^{1-\alpha} - F. \tag{4.12}$$

This function is similar to (4.2): y_i is the production of a firm in sector i.[9] It uses labor and bundles of intermediate product $Q_{i,k}$ with $k = 1, 2$ the supplying sector. The bundles are from formula (4.11); we again assume that the elasticity of substitution between products of different producers in the same sector, σ, is the same for final and intermediate demand. Knowing this, when it comes to pricesetting the producer does not have to worry about the different clients and can set the same price for all: the usual markup over marginal costs, MC $\cdot\, \sigma/(\sigma - 1)$. Finally, each firm faces a fixed cost F that is paid in the final product.

There are a number of constants in formula (4.12). The θ's are defined as follows:

$$\begin{aligned} \theta_\alpha &= \alpha^{-\alpha} \cdot (1-\alpha)^{\alpha-1}, \\ \theta_{\eta,i} &= \eta_i^{-\eta_i} \cdot (1-\eta_i)^{\eta_i-1}, \end{aligned}$$

and serve to normalize associated costs. The marginal costs for a firm in sector i in the *home* region are

$$\begin{aligned} MC_i &= \phi \cdot w^\alpha \cdot (\tilde{G}_i)^{1-\alpha}, && (4.13) \\ \tilde{G}_i &= \psi_i \cdot G_1^{\eta_i} \cdot G_2^{1-\eta_i}, && (4.14) \\ G_i &= \left(n p_i^{1-\sigma} + n^* \left(\frac{p_i^*}{\tau} \right)^{1-\sigma} \right)^{\frac{1}{1-\sigma}}. && (4.15) \end{aligned}$$

---

[9] Notice that we do not index by region—it is assumed that the production functions are similar in both regions.

These expressions follow directly from (4.12). Note that G_i is the price index of goods from sector i in the *home* region, and $\tilde{G}_i$ is the price index of intermediate goods, *used by sector i*, in the *home* region. For the *foreign* region, G_i^* and $\tilde{G}_i^*$ could be defined.

From (4.13)-(4.15), we see that marginal costs are a weighted geometrical average of wage costs in the *home* region and the price of intermediate goods. The latter consist in turn of a weighted average of the price indices of goods from sectors 1 and 2. The price index of goods from sector i, finally, is a weighted average of all prices in the sector, both of firms in the *home* region and in the *foreign* region. Note that in order to use prices from the other region, we have to take the transport costs into account.

From the coefficients η_i in formula (4.12)we can construct a two-by-two IO matrix,

$$IO = \begin{bmatrix} \eta_1 & \eta_2 \\ 1-\eta_1 & 1-\eta_2 \end{bmatrix}. \tag{4.16}$$

We defined an IO matrix as containing the shares of the budget for intermediates that go to the different sectors. The columns sum to one, indicating that the total budget for intermediates is exhausted. This matrix can be constructed from an IO table by dividing the entries (the flow of trade from one sector to another) by their column sums. Thus, for instance, η_1 is the share of their budget for intermediate goods that firms in sector 1 spend on products from their own sector.

So far, we have left the scaling constants ϕ and ψ unspecified. As before, we will set $\phi = 1/\theta_\alpha$ for a simplification of (4.12). For the same reason, we could set $\psi_i = 1/\theta_{\eta,i}$ for all i for a baseline result and compare cases where different values of ψ lead to a different outcome. However, another issue comes to the front, which is the result of our assumption that discrete sectors are aggregated using a Cobb-Douglas function (as in 4.10). First of all, we need to recognize that the precise categorization introduced into the model is often the result of a judgement call, affected by factors such as data availability. Depending on data, we could carve the economy into two sectors or into twenty. If we want to compare the results of the two-sector model to those in a twenty-sector model, they should at least be the same for some special twenty-sector cases.

A natural point of departure is a generalization of the one-sector model of section 4.2 into an s-sector model ($s \geq 2$) with the IO matrix $\eta_{i,j} = 1/s$ for all i, j. In this case, intermediate bundles consist of equal amounts of products from *each* producer, regardless of the number of sectors. It seems natural that the price index of intermediates $\tilde{G}$ (which is equal across sectors) should be the same for any value of s. Whether the intermediate goods come from two sectors or from twenty seems more of an administrative concern than something that would influence the price of that bundle.

In order to achieve such equivalence, it is important to remember that the monopolistic competition that exists between firms does not cross the sector boundary. Firms within the same sector are monopolistic competitors, but each sector in total is guaranteed a fixed share of each budget spent on industrial goods. The latter is a result of our various Cobb-Douglas assumptions. Therefore, the love-of-variety effect that causes increasing returns to the number of firms, only works on the number of firms *within a sector*. In this model, carving the economy up into many sectors thus has the undesirable effect of reducing overall efficiency.[10]

This can be seen as follows: suppose we convert a single-sector economy into an s-sector economy, with all entries of the IO matrix equal to $1/s$. The price index of intermediate goods in each case is

$$\begin{aligned} \tilde{G} &= \left[n \cdot p^{1-\sigma}\right]^{\frac{1}{1-\sigma}} \\ &= n^{\frac{1}{1-\sigma}} \cdot p \qquad (4.17) \\ \tilde{G}_s &= \left[\frac{n}{s} \cdot p^{1-\sigma}\right]^{\frac{1}{1-\sigma}} \\ &= \tilde{G} \cdot \left[\frac{1}{s}\right]^{\frac{1}{1-\sigma}} \qquad (4.18) \end{aligned}$$

Here, $\tilde{G}$ is the one-sector price index and $\tilde{G}_s$ is the s-sector price index of a bundle of intermediates. As s is larger than one, $\tilde{G} < \tilde{G}_s$, or, the choice of the number of sectors s affects the price of intermediates. In order to preclude this result we introduce s, the number of sectors, into ψ_i: below, we set

$$\psi_i = \frac{s^{\frac{1}{1-\sigma}}}{\theta_{\eta,i}}. \qquad (4.19)$$

This way, results between different categorizations are comparable in principle.

The model is now operational. We can compute the equilibrium for a region, given the number of firms and the prices of goods in the other region, and given the demand from the other region for home products. A detailed description of the solution method is provided in the next paragraph.

4.2.4 Solving the model

We arrive at the numerical solution of our model in different stages. The order in which these stages are computed reflects our assumptions about the speed of movement: because we think that workers shift sectors faster than firms can enter or exit, we solve the model in this order, starting with

---

[10] In effect, we have assumed that there exists a large payoff to variety within a sector but not between them. An alternative approach, in which this problem does not occur, was discussed in the previous chapter.

an initial guess for N, the numbers of firms and and L, the sector-specific numbers of workers per firm:

- Given N and L, we compute regional wages, prices of intermediates and prices for each type of product. The latter two have to be solved simultaneously.
- Given wages and prices, we compute the demand for each type of good. This leads to profits, which are distributed among workers in the same region as the firm (we assume local ownership). Profits again lead to extra demand, so the two have to be solved simultaneously.
- With demand and supply for each firm thus computed, we change L to correct any imbalances between supply and demand. This brings us back to the first bullet point. L is changed iteratively in a numerical Gauss-Newton procedure.
- Now that demand equals supply and wages balance the labor market, we look at profits. Positive profits lead to entry, losses lead to exit. This changes N and brings us back, again, to the first bullet point.
- The routine stops when profits or the corresponding number of firms are zero. For N as well, a Gauss-Newton procedure is used.

We add restrictions. By indicating which sector-region combinations should have no firms, or equal numbers of firms, we study the different equilibria.

4.3 Types of solutions

4.3.1 A Taxonomy

We now look at possible equilibrium results of our model. The results depend very much on the share of industrial products in consumption, $\mu_1 + \mu_2$. If it is larger than $1/2$, the manufacturing sector has to be spread over two locations. If it is smaller than $1/2$, the sector can agglomerate into one region. When we assume that $\beta < 1$, so that there are decreasing returns in agriculture, four types of equilibrium can occur:

1. All industrial activity agglomerates in one region. This happens when $\mu_1 + \mu_2 < 1/2$, the elasticity of substitution σ is small and firms from different sectors use each other's products as intermediary inputs. The region without industry imports industrial products for final consumption, and trades them for agricultural products. In the figures below, this solution is indicated as DEV (for a *developed* region versus an undeveloped region.) The number of firms in each sector is

determined by demand, both final and intermediate. This means that the sectors need not have the same number of firms, as they do in the figure.

2. Both regions get exactly the same, positive number of firms in each sector. This happens when the share of industrial products in consumption is high, and firms from different sectors use each other's products as intermediates. There is only intra-industry trade in equilibrium, with no trade in agricultural products. The values of μ_i do not matter. In the figures below this solution is indicated as SYM, indicating a symmetric equilibrium. Note that the sectors are not necessarily symmetric to each other, as the number of firms is once again determined by intermediate and final demand.

3. Each region gets only one of the two industries. This happens when firms in one sector use very little of the other sector's products as intermediate inputs and trade costs are reasonably low. In equilibrium, there is only inter-industry trade. This solution is indicated as AGL, the outcome where each industry agglomerates. The number of firms in each region depends on the demand for its sector.

4. Finally, in the fourth type of equilibrium, one region specializes completely in a particular sector, say sector 2. The firms from sector 1 agglomerate in the other region, but demand for sector 2 goods is so large that it cannot be filled by the firms in the specialized region. Thus, the some sector 2-firms also appear in the other region as well. The relative number of firms in each region depends on the size of demand, but the sector that has firms in both regions is unevenly distributed. This solution occurs only when $\mu_1 + \mu_2 > 1/2$, and is indicated as OVF, for overflow.

Equilibria DEV and SYM are the results that Venables (1996a) found. For intermediate values of transport costs, all of industry agglomerates into one region. This is equilibrium DEV. For extremely low and high values of transport costs, equilibrium SYM obtains.

The AGL equilibrium was found by Krugman and Venables (1996), in a model without agriculture. The separation of industries occurs when the links between the different sectors are weak, but they are strong within the sectors. The equilibrium obtains under small values of the costs of transport.

The fourth type equilibrium (OVF) has, to our knowledge, not been discussed in the literature. It is an interesting type, where most of industry agglomerates into one region, but a small number of firms finds it profitable to settle in the agricultural region. It bears resemblance to the equilibrium discussed in figure 4.8, in the one-sector case.

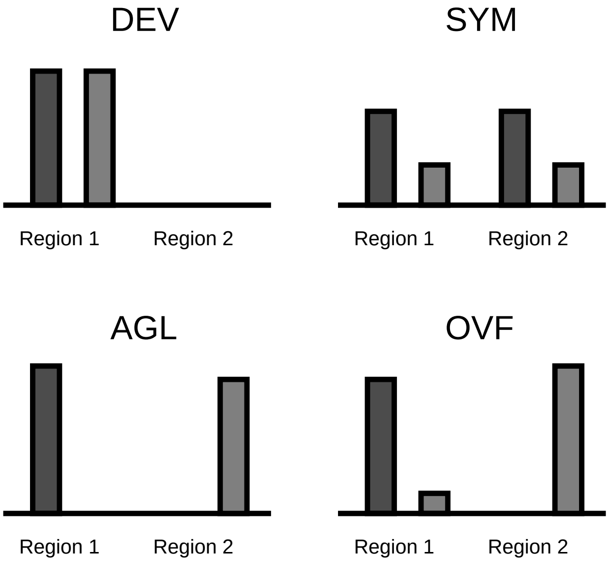

Figure 4.9: The four types of equilibrium, stylized. We look at two sectors, two regions. The graphs show the two regions left and right, and the number of firms in the two sectors is indicated by the height of the two bars.

Figure 4.9 shows a stylized impression of the four types of equilibrium for future reference.

4.3.2 Stability of equilibria

Given that the initial conditions play a role in the final equilibrium of the model, we have to characterize the results of a particular set of parameters in terms of stability, as we did before. This means that we cannot predict whether a certain IO matrix will produce a separation between the industries or a mix of both sectors in both regions.

The parameters of the model are chosen with care. We take the elasticity of substitution $\sigma = 7$, a rather low value that is often chosen because it makes it hard for industries to substitute between different intermediate inputs. This way, the linkages are strong and we are able to see some interesting results. The size of the manufacturing sector, $\mu_1 + \mu_2$ is taken equal to 0.9, so that manufacturing is too big to fit in one region (we look at the other case below). This way, we cannot find the total agglomeration result (type DEV). The other parameter values are fixed costs $F = .008$, labor share $\alpha = 0.6$ and the parameter that governs agricultural decreasing returns, $\beta = 0.9$. Transport costs, finally, are set to $\tau = 0.9$.

The results of this simulation are in figure 4.10. These are the stability characteristics for each possible IO matrix. The figure is laid out as follows: the value of η_1 (see formula 4.16) is on the horizontal axis and the value

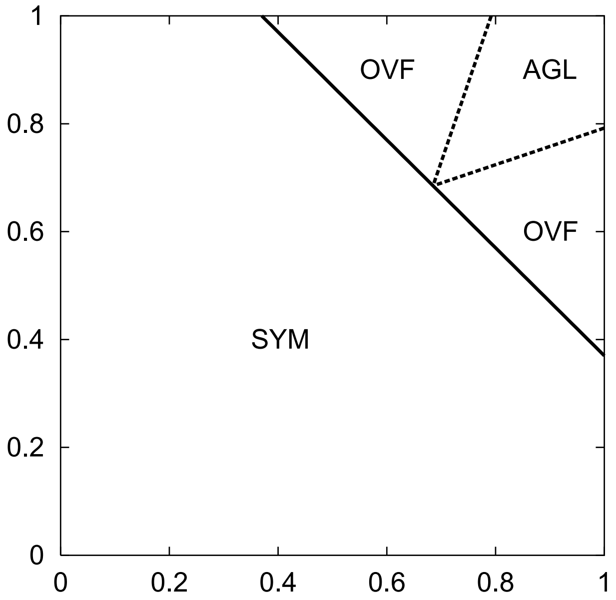

Figure 4.10: Each point in this square represents a 2-by-2 IO-matrix. The upper-left entry of the matrix (the share of sector-1 products in sector 1′s intermediate good) is on the vertical axis. On the horizontal axis is the bottom-right entry (the share of sector-2 products in sector 2′s intermediate good). Because these two entries automatically define the remaining two, the square represents all possible 2-by-2 IO matrices. There are three possible equilibria in the model that correspond to the indicated areas in this figure. SYM is the symmetric equilibrium and AGL is a complete separation of the two sectors whereby each agglomerates in its own region. Finally, OVF is the so-called ‘overflow’ equilibrium.

of $1 - \eta_2$ is on the vertical axis. Any point on the square thus defines a certain IO matrix, and the whole square represents the set of all possible IO matrices.

We indicate where each of the four different types of equilibrium is stable, using the abbreviations introduced above. As it turns out, the areas of stability are mutually exclusive. This can be seen in figure 4.11, which is similar to figure 4.3 in the sense that on the vertical axis, we see the maximum eigenvalue and the profit of a breakaway firm. On the horizontal axis, however, are not transport costs τ (those are constant) but is η_1, which for this graph is equal to $1 - \eta_2$. That is, the graphs shows what happens along a diagonal line in figure 4.10

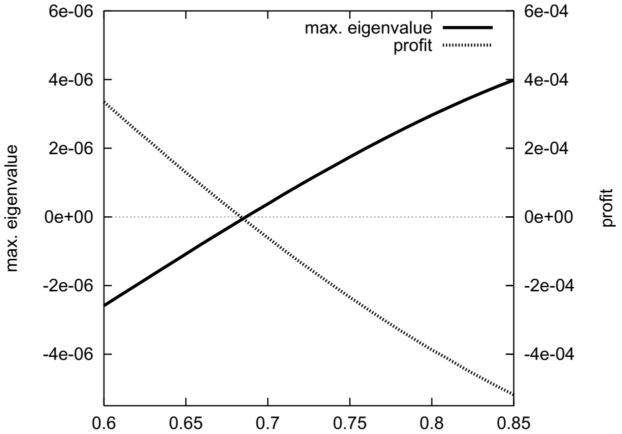

Figure 4.11: A trip along the diagonal of the 'map figure'

Is this pattern of stability typical for all parameter values in this model? We present another outcome in figure 4.12, where the parameters have been changed to $\sigma = 4$, $\tau = 0.7$, $\alpha = 0.37$ and $F = 0.001$. These are the same as in the one-sector model that was discussed using figure 4.3. Notice that the OVF equilibrium no longer obtains. The equilibrium in which two groups of firms each agglomerate in their own region is stable for a wider range of parameters. Elsewhere, we find a symmetric outcome.

4.4 Conclusions

In this chapter, we again looked at a Venables-type economic geography model where the industrial firms are divided up into different sectors. In a departure from the assumptions of chapter 3, we opted for a discrete number of sectors. Therefore, we can identify groups of firms who share the same characteristics with respect to their intermediate goods preferences.

How will these groups of firms spread out over different regions, given

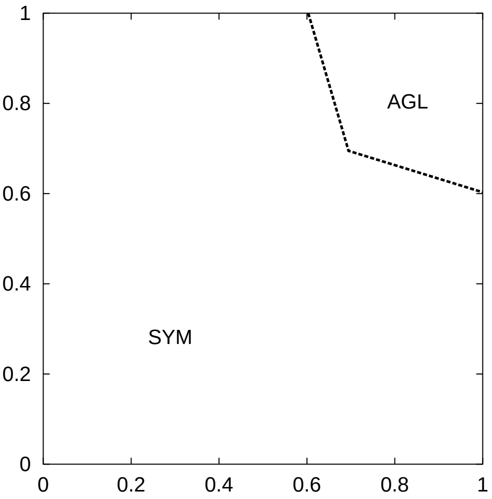

Figure 4.12: Each point in this square represents a 2-by-2 IO-matrix. The upper-left entry of the matrix (the share of sector-1 products in sector 1's intermediate good) is on the vertical axis. On the horizontal axis is the bottom-right entry (the share of sector-2 products in sector 2's intermediate good). In this case, $\sigma = 4$, $\tau = 0.7$, $\alpha = 0.37$ and $F = 0.001$, as in figure 4.3.

that there are costs of transport between the two regions? In chapter 3, we saw that simple cases yield simple answers: when two groups of firms only use each other's products as intermediate inputs, these groups will each cluster in a different region. That way, they profit from each other's closeness while staying away from the negative externalities of other firms, such as a higher wage rate.

In this chapter, we were able to take this model further, looking at the all the possible degrees to which to sectors can be intertwined. That is, we looked at all possible input-output matrices between two sectors and mapped the type of spatial equilibrium that obtained on the space of IO-structures.

The simple cases of chapter 3 can be reproduced in this setting: we see them in the right-top and left-bottom corners of figure 4.10. The right-top corner is the case where two groups of firms only use inputs from their own group. Each group will agglomerate in one region. In the left-bottom case firms in each group only use output from the *other* group as intermediates. In this case, each region will have a 50-50 mixture of firms from each of the two groups.

What is interesting is the middle ground. The rest of figure 4.10 shows how the borders between the different equilibria lie. We see (figure 4.11) that there is a clear point where the stability of the different equilibria switches. This means that a small change in technology, whereby a parameter of the IO matrix changes, can cause a dramatic change in the type of equilibrium that obtains. We have seen a similar disproportionate effect of a changing parameter before: a small change in transport costs can switch the model's equilibrium as well (see figure 4.3, for instance).

This chapter has thus demonstrated the dependence of model outcomes on IO parameters, and shown that a small change in these parameters can change the outcome dramatically. But it has also shown a useful way of computing the equilibrium of a model with a number of discrete sectors, when those sectors have ties of different strength, summarized in the IO matrix. If we can find relevant parameters for such a model, we can use it to compute counterfactuals. This would make the model into a tool for policy evaluation. We start work on finding the relevant parameters in chapter 5 below. Using the results of that chapter, we will construct a model of the Dutch economy in chapter 6, in which there are fourteen different sectors. This model will then be used in a policy evaluation exercise.

Chapter 5

Estimation of parameters in an economic geography model

5.1 Introduction

In the previous chapters, we have seen how models of economic geography explain the agglomeration of economic activities. The explanation involves increasing returns to scale at the firm- and industrial level. The models show that if trade is impeded by costs of transport, it is possible, depending on the parameters, that a concentration of activity arises as a result of economic mechanisms.

But while such concentrations of economic activity surely exist in practice, as many real-world examples will testify, it may be difficult to determine if the purported mechanisms of economic geography theory have anything to do with them. If we are to show that this is the case, we need to distinguish this cause for agglomeration from other possible explanations, such as natural geographical circumstances. Ellison and Glaeser (1999) claim that natural advantages, such as the presence of a natural harbor or a particular climate, can be used to explain "at least half of observed geographic concentration" (p. 316). Setting these causes, which are often called 'first nature,' apart from economic incentives to agglomerate (or, 'second nature') is a methodological challenge.

In this chapter, we will survey the empirical literature on economic geography. Furthermore, we will attempt to estimate the parameters of the model of the previous chapter, using data on shipments between American states. The payoff of this exercise will be twofold: not only will we find out about the relevance of our theoretical exercises, but the attempt to fit the theory of economic geography to real-world data will also give us a calibrated version of our model.

Firstly, the estimation of model parameters serves the purpose of validating our model as an explanation of events in the real world. The model's performance, its goodness of fit, will tell us if we should not stick with a simpler alternative. As an added advantage, the estimated parameters can be interpreted as variables of economic interest in their own right. As we have seen in earlier chapters, parameters such as the elasticity of substitution between differentiated goods and the costs of transport, figure in models of economic geography. An estimated value of these parameters would be interesting, even outside the context of a model.

Secondly, an economic geography model calibrated to real-world data would be a useful tool for many purposes. Models from this class, because of their explicit microeconomic foundations, lend themselves readily for the evaluation of counterfactuals. As such, they can be a tool to evaluate proposed economic policies in the light of their geographical repercussions. For instance, questions as to how the construction of a new road would affect the regions around it can be addressed using a calibrated economic geography model. Not only can we compute the new equilibrium after a change has taken place, but we can also assess the resulting change in welfare. There exists a strand of literature that employs these models for precisely this purpose, which includes Venables and Gasiorek (1996) and Bröcker (1995, 1999).

Another interesting phenomenon that could be analyzed with a calibrated model is the spatial distribution of the effects of a local shock in demand. It is assumed that such a shock will have repercussions outside the originating region, but exactly how far is unknown. We look at the analysis of this question by Hanson (1999), who analyzes the spread of a shock in demand from the center of the United States. We will evaluate a similar shock with our calibrated model below.

After the survey of the empirical literature in section 5.2, we use the rest of the chapter for an attempt to estimate the parameters of an economic geography model. The model is drawn up along the lines of chapter 4, using vertical relations between firms as the agglomerating force. In section 5.3, we discuss the model that will be used in this estimation. As before, the model explains flows of trade between different regions by relating them to the number of firms and the number of consumers in each region. The number of firms in the exporting region has a positive effect on the level of trade, because of the Armington assumption: each extra firm offers a unique product which is desired in all regions. Firms and consumers in the importing region both exercise demand, which increases when there are more of either group. Because the model also involves transport costs that increase with distance, it predicts large trade flows between states that are close. Taken together, this specification is reminiscent of the gravity trade model (by, for instance, Tinbergen 1962). Named for its mathematical analogy to that well-known force of attraction, the model predicts that trade is proportional to the size of the trading economies, and is inversely proportional to their distance. While its origins are certainly agnostic, it has been derived in models of monopolistic competition and Deardorff (1995) shows that it can also be derived in a neoclassical framework.

But while we find that our model resembles the gravity model of trade in its specification of the trade equation, it goes beyond gravity when we take into account its general equilibrium nature. Where the gravity model only predicts the flow of trade, given the economic size of regions and their distance, the economic geography model explains trade, prices *and* economic size in one consistent framework. We will use a gravity-like equation in the first part of our estimations, but will move beyond the results to verify the coherence between wages, prices and market access for the different regions.

Our data is introduced in section 5.4. It consists of the 1997 US Commodity Flow Survey, which contains measurements of the bilateral shipments between the 51 American states. The Commodity Flow Survey is a relatively underexploited dataset. It is based on a survey conducted by the Bureau of Transportation which asks for origin and destination addresses, value and weight. The 1993 version of the survey was used by Wolf (1997) to estimate gravity equations and examine the home market effect in US states.[1]

Estimation is done in Section 5.5. We proceed in two different ways. Section 5.5.1 discusses an attempt to estimate the parameters of the model using the methods of Redding and Venables (2001); section 5.5.2 applies a different method to estimate the same model from the same dataset. The latter method leaves us with a calibrated model, that can be used for the analysis of counterfactuals. We use it to compute the spatial effect of a shock in wages in the United States, and experiment with the lowering of trade costs between regions. Finally, Section 5.7 concludes.

5.2 Estimation in the literature

Overman et al. (2001) survey the empirical literature concerning economic geography. Their survey is built around a canonical theoretical model, similar to the models developed in previous chapters. Different aspects of the model are used in various exercises, all which serve the purpose of testing the model's predictions.

The analyses that are surveyed can be divided into three areas: firstly, measurements of the importance of transport costs are introduced, in order to show that distance matters for economic activity. Transport costs are defined broadly, including not only the price of actual transport but also the time in transit and costs of information. Besides measuring such costs directly, by comparing *fob* and *cif* prices, an indication of the role of distance

---

[1] The home market effect is a result in trade theory that follows from monopolistic competition. It was discussed in section 2.B.

is given by gravity estimations. The fact that a region's trade with similar partners declines with their distance shows that transport costs indeed form a burden.

The second area of research focuses on the fact that the distribution of economic activity is usually uneven across regions. This can be shown directly, for instance with the use of Gini-coefficients. Though it is in accordance with economic geography theory, the fact that activity is distributed unevenly can be explained by other factors as well. As such, it cannot be the only evidence for the theory. However, the uneven structure of production can be used to test for the home market effect, which does point to an economic explanation. We look more closely at this line of research below, in section 5.2.1.

Finally, a strand of literature exists that relates a region's prices of immobile factors to the degree of access to other regions. The idea is that remote locations, facing high prices for intermediate goods and a bad export position, are forced to pay less for immobile factors of production. Specifically, attempts to relate the spatial wage structure to market access have been successful. We discuss this line of literature below, in section 5.2.2.

After our discussion of econometric estimation, we take a look at the more practical art of model calibration in section 5.2.3. In the papers discussed in this section, the researchers have built an economic geography model whose virtues are not under discussion. These models need to be calibrated so that initially, they are able to replicate the data. In the process of fitting the model, interesting choices have to be made on certain parameter values.

5.2.1 Measuring the home market effect

An important series of articles on the matter of testing for economic geography are Davis and Weinstein (1998a, 1998b, 1999). They use the home market effect discussed in section 2.3.3 above to put a trade model to the test. The model uses the Heckscher-Ohlin theory at the level of industries, and allows for a number of alternatives at the level of individual goods.

The home market effect may be observed if a region or a country has a large idiosyncratic demand for a particular good. In the Heckscher-Ohlin theory of trade, where the production structure is driven by factor endowments, such a region will in general be an importer of that good. The presence of an extraordinary level of demand does not affect the location of production. Even though local producers will satisfy the demand to some extent, they will not cover all of it, hence the importer status of the region.

In a world where producers only use one factor and compete within an MC framework, things are different, and the home market effect surfaces: the region with the large demand for a specific good will be a net *exporter* of that good. The producers, rather than being driven by factor scarcity, realize that they are best off producing in only one location because of returns to scale, and prefer the region with the largest demand because of transport costs. Other regions are then serviced from this one, leading to the net export result. In terms of the model, a large demand component is matched more than one for one by the region's production. This is called the home market effect; it is discussed in Section 2.B.

In the empirical specification, the authors lump together several goods into an 'industry,' explain industry location by endowments and look how the production of particular goods within the allotted industry production is distributed. In Davis and Weinstein (1998a), the data come from the OECD and concern national manufacturing production of the member countries. Results are meagre, in the sense that most trade can be explained by the traditional Heckscher-Ohlin model. When the model is estimated with data for regions in Japan (Davis and Weinstein 1998b), the results are quite different. Estimated for each of the industries separately, two out of six feature a marked 'home market effect.'[2] The authors indicate that this could be the result of lower transport costs and greater factor mobility at the regional level.

The tests by Davis and Weinstein must be seen as a first coarse investigation into the relevance of economic geography models. Many assumptions are made: at what level of aggregation do 'industries' stop, and do we observe 'goods' where the market is monopolistically competitive? What goods form an industry together? How do arbitrary definitions of regions and groups of goods affect the results?

Two further objections to this strand of research are raised by Brakman et al. (2001, section 5.4.3). First of all, the home market effect that is tested by Davis and Weinstein does not exclusively identify models of economic geography. Brakman et al. argue that the effect is also present in other models that do not have the crucial endogenous market size. Secondly, the home market effect is not very robust in models of economic geography—there exist variants that do not feature the effect. This renders testing on the basis of it inconclusive at best.

5.2.2 Estimating a spatial wage structure

A result that can be obtained using models of Economic Geography, but does not appear in rival models, is the negative relationship between a region's wage and its distance from other economic activity (Brakman et al. 2001, section 5.5). Measurement of this relationship would be a clear indication of the relevance of Economic Geography models, as well as give insight into the values of it key parameters.

---

[2] These industries include: transportation equipment, general machinery, electrical machinery and precision instruments. Indeed these skill-intensive goods seem to be among the ones where differentiation is possible, as opposed to manufactured bulk goods.

The relationship can easily be derived from the expression for a firm's foreign demand in formula (2.8). If we generalize from a model with two regions to a model with R regions, we can write demand for firm i in region j as

$$x_{i,j} = p_{i,j}^{-\sigma} \sum_{s=1}^{R} E_s q_s^{\sigma-1} \tau_{s,j}^{\sigma-1}$$

where $\tau_{s,j}$ is the iceberg-transport cost parameter between regions s and j. It is often parametrized as $\exp(-\xi d_{s,j})$ where $d_{s,j}$ is the distance between the two locations.

Knowing that $x_{i,j}$ must be equal to its optimal level because of the free-entry assumption, and knowing that $p_{i,j}$ is usually just a multiple of the local wage level w_j, we can rewrite the above as

$$w_j = \left[\sum_{s=1}^{R} E_s q_s^{\sigma-1} \exp(-\xi d_{s,j})^{\sigma-1} \right]^{1/\sigma} \tag{5.1}$$

Hanson (1998, 1999) notices that this spatial wage structure, predicted by Krugman's (1991a) model, resembles the market potential function of Harris (1954). He estimates the spatial wage structure based on a market potential formulation, and goes on to estimate the reduced form of Helpman's (1998) adaptation of the Economic Geography model, which is similar to (5.1).

The estimation in Hanson (1999) is done using a first-difference form of (5.1) to account for the (unchanging) external qualities of the land. Hanson uses data for 3,075 U.S. counties and finds the expected sign for all coefficients, all of which are statistically significant. Using these coefficients, he is able to simulate the effects of a shock in demand in Chicago for the rest of the continental United States.

Brakman et al. (2001) criticize Hanson for failing to recognize the importance of international trade in his estimations. They offer an alternative study into the spatial wage structure in East and West Germany. Their estimates are similar to Hanson's (1998), but do not change very much when trade with the rest of Europe is factored into the model.

We see that two methods have been used to estimate the parameters of economic geography models using statistical methods. The works of Hanson (1999) and Brakman et al. (2000) use the wage structure of US counties and German districts, respectively, to parametrize a Krugman-type model. Davis and Weinstein estimate the coefficients of a model of international trade, interpreting the values in the light of monopolistic competition theory.

In section 5.5, we will discuss the estimation of a Venables-type model, where links between industries work through the use of intermediate goods.

We delay this discussion because it makes use of the model that is to be introduced in section 5.3 below. We will apply the methodology, introduced by Redding and Venables (2001), to a new set of data.

In the next paragraph, we look at the methods that have been used to calibrate models of Economic Geography to data. While this is not estimation *per se*, since the amount of data used is too small to disagree with the model, some insights carry over to the methodology of section 5.5.

5.2.3 Model calibration

Venables and Gasiorek (1996) use a model in which industries are linked through intermediate inputs. They apply this model to several European regions, including the Iberian peninsula and Greece. Factor shares and industry sizes are taken directly from national accounts data. Elasticities are presumably specified using estimates from the literature, but their values are not reported.

Transport costs are a function of the distance between two regions. This function is approximated iteratively: starting with an initial guess, the authors solve the model and compare (gross) trade between regions with actual data. From this comparison, the function that maps distance to transport costs is re-specified, until the outcome matches the data.

Once the gross streams of trade are fitted, the authors go on to specify sector-specific transport costs-functions, preserving size of the total streams.

Bröcker (1998) specifies a general equilibrium model that is reminiscent of the simple Venables model of chapter 4. There are some differences, however: the workers in the model do not choose their sector endogenously, and it is assumed that their wages are equal. The author finds that flows of trade between regions vary inversely with the distance between the trading partners and a coefficient $\rho = \sigma \cdot \zeta$. Here, σ is the elasticity of substitution and ζ a coefficient related to the costs of transport. Bröcker is able to estimate ρ from trade data, but then has to separate out σ and ζ in order to specify his model. He accomplishes this by looking at the share of transport costs in the total value of trade, knowing from his model that this number is equal to $1/2\sigma$. Data on transport costs thus imply that σ is smaller than 10. Furthermore, we know that the markup factor is equal to $\sigma/(\sigma - 1)$. Data on markups suggest that the value of σ must be larger than 5. Finally, the author finds that the outcome of the model changes very little for σ's between 5 and 10.

5.3 Vertically linked industries and agglomeration

In the rest of this chapter, we will develop a Venables-type of economic geography model and estimate its parameters using data on trade between

American states.

As before, we use the structure that was introduced by Venables (1996a) and that was discussed in the previous chapters. In this model, connections between firms exist because of the use of each other's intermediates in production. We previously studied how the input-output relations between sectors influenced the possible equilibria of this model. In this chapter we assume, rather implausibly, that each producer uses *all* available products as intermediate inputs. In this case, we do not need to specify different sectors, so that we have a single industrial sector that works under Monopolistic Competition (MC) and IO-matrix is collapsed to the number 1. The single-sector assumption is an obvious shortcut with which we will live in this chapter, as it facilitates estimation; we will estimate a more general version in chapter 6.

Next to the industrial sector, whose products can be traded across regions, we assume the existence of a 'local products' sector, whose production is nontradeable. This sector comprises activities such as local services and the production of locally consumed goods. The local sector is perfectly competitive and uses a simple linear production technology where the only input is labor.

In the industrial sector, firms use labor and products from other industrial firms as inputs. Each firm i in region r produces a single variety with production function

$$Y_i = \theta_\alpha L_i^\alpha Q_i^{1-\alpha} - F$$

where $\theta_\alpha = \alpha^{-\alpha}(1-\alpha)^{\alpha-1}$ is a constant scaling term, L_i is the amount of labor employed and Q_i an aggregate of intermediate products used. Finally, a fixed cost F is paid in the company's own product. Adding a fixed cost excludes the solution where a firm would produce an infinite number of goods in infinitely small amounts to exploit the increasing returns to variety.

For the aggregation of industrial products into a single intermediate product we employ the usual CES function, so that

$$Q_i = Q\left(x_1, \ldots, x_N\right) = \left[\sum_{j=1}^{N} x_j^{1-\frac{1}{\sigma}}\right]^{1/\left(1-\frac{1}{\sigma}\right)} . \tag{5.2}$$

As discussed, the intermediate product Q is composed of products of all N firms in the economy. Again, it pays to be close to other firms for two reasons: the lower price of intermediates and the local demand for your own product (these are the forward and backward linkages of Hirschman 1958).

Given that each industrial firm i in region r faces exactly the same conditions, they will all demand the same amount of intermediate product, which we denote Q_r. The price index associated with Q_r is

$$\begin{aligned} G_r = G\left(p_1^r, \ldots, p_N^r\right) &= \left[\sum_{j=1}^{N} \left(p_j^r\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}, && (5.3) \\ &= \left[\sum_{s=1}^{R} n_s \left(p_s^r\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}} && (5.4) \end{aligned}$$

where p_j^r is the price of firm j's products in region r. Given that all firms in the same region use the same price, we arrive at (5.4), with p_s^r the price of an industrial product from region s in region r ($1 \leq r, s \leq R$) and n_s the number of firms in s. Prices of the same product differ between regions, as there are different costs of transport. The exact nature of the transport costs is discussed below.

Consumers optimize a simple Cobb-Douglas utility function over local and tradeable goods. In region r, the function is

$$U(Z, x_1, \ldots, x_N) = U(Z, Q(x_1, \ldots, x_N)) = Q^{\mu_r} \cdot Z^{1-\mu_r} \qquad (5.5)$$

Here, Z is the quantity of local goods consumed. Industrial products enter as multiples of the aggregate Q, defined in (5.2). This implies that consumers and producers have the same elasticity of substitution between products, equal to σ. Because of this, we can invoke the result that all firms in a region use the same, optimal price, for final and intermediate demand. The price is a markup over marginal costs:

$$p_r^* = \frac{\sigma}{\sigma - 1} w_r^{\alpha} G_r^{1-\alpha}. \qquad (5.6)$$

We denote the number of people working in region r as L_r. Since a fraction $1 - \mu_r$ of the region's wage income[3] goes to local producers (per formula 5.5), it follows that a fraction $1 - \mu_r$ of all workers are active in the local-goods sector. The remaining $\mu_r L_r$ people work in the industrial sector. From the size of the workforce we can compute the number of firms, given free entry and exit so that each firm makes a profit of zero. For then we must have that for any firm i, wholesale profits exactly compensate the fixed costs that were incurred, or

$$\begin{aligned} Y_i &= (\sigma - 1)F \quad \Rightarrow \\ \sigma F &= \theta_{\alpha} L_i^{\alpha} Q_i^{1-\alpha} \\ &= \frac{L_i (w_r / G_r)^{1-\alpha}}{\alpha} \end{aligned}$$

---

[3] In this static model, wage income is the only income since we abstract from saving and capital.

where we used the definition of θ_α, the markup equation (5.6) and the fact that, after optimization, $Q_i = \frac{(1-\alpha)w_r}{\alpha G_r} L_i$. This gives us the optimal amount of labor used by firm i in region r, L_i^*. The number of firms in region r can then be computed by dividing the number of industrial workers in the region by L_i^*.

$$n_r = \frac{\mu_r L_r (w_r/G_r)^{1-\alpha}}{\alpha \sigma F} \tag{5.7}$$

The number of firms in a region varies as the local price index G_r changes.

Transport costs are incorporated in the model using the 'iceberg assumption', whereby transport charges are incurred in the product itself. The amount that needs to be shipped to get one unit of the product to arrive from location s in location r, T_{sr}, corresponds to the distance travelled as

$$T_{sr} = \exp\left(\tau d_{sr}\right)$$

where τ is a positive parameter. Alternatively, one could consider T_{sr} a markup over the home price: for each region s, there holds $p_s^r = p_r^r T_{sr}$. We rewrite the price index in (5.4) as

$$G_r = \left[\sum_{s=1}^{N} n_s \left(p_s T_{sr}\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}. \tag{5.8}$$

Given parameters, wages, and numbers of workers L_r we can now solve the model in terms of prices, price indices and the equilibrium number of firms. These three sets of region-specific variables p_r, G_r and n_r can be stored in three vectors of length R. Each is defined in terms of the other: we have vector $\mathbf{p}$ from equation (5.6), $\mathbf{n}$ from equation (5.7), and $\mathbf{G}$ from equation (5.8). Ideally, we would solve this system of $3 \times R$ equations analytically. However, this is not possible for reasons that were discussed in chapter 2, so we have to rely on numerical methods instead. We use an iterative routine in Matlab to find the values of $\mathbf{p}$, $\mathbf{G}$ and $\mathbf{n}$ that satisfy the equations. In general, this routine finds the equilibrium very quickly.

In this model, a region's expenditures on industrial products comes from final and intermediate demand. In region r, the L_r consumers spend $E_r^f = \mu_r w_r L_r$ on industrial products from all over the economy. Firms in region r buy intermediates, spending an amount directly proportional to their total wage bill: $E_r^{\text{int}} = \mu_r L_r w_r \left((1-\alpha)/\alpha\right)$ (this follows from the Cobb-Douglas production function). The total expenditure in region r is thus $E_r = E_r^f + E_r^{\text{int}} = \mu_r L_r w_r / \alpha$.

Depending on the parameters, this model can have several different equilibria. If the costs of transportation and the share of intermediate products in production are low, and the elasticity of substitution σ is high, then economic production will be distributed proportional to the population size. But if transport costs are high, intermediate products are important and σ is low, then production can agglomerate in one or a few regions (For a full derivation of these results, see Venables 1996a). Which region gets the agglomeration is decided by initial conditions.

Region r's expenditures on industrial goods, E_r, are spread over all industrial producers in the economy. Products from region s cost $p_s^s T_{sr}$ and the price index is as in (5.8). Standard Dixit-Stiglitz optimization leads to the familiar result that in region r, the demand for a product from region s is

$$D_{rs} = E_r \left(p_s^s T_{sr}\right)^{-\sigma} G_r^{\sigma-1}. \tag{5.9}$$

Given that there are n_s firms in region s, each producing a unique differentiated product with the same price, the total demand in region r for products from region s is n_s times the expression in (5.9). To get the value of this stream of goods, we also multiply by the price[4] which gives

$$X_{rs} = n_s \left(p_s^s\right)^{1-\sigma} T_{sr}^{1-\sigma} E_r G_r^{\sigma-1} \tag{5.10}$$

with X_{rs} the value of shipments from s to r.

Of course, equation (5.10) is reminiscent of the gravity equation. The term n_s is indicative of the economic size of the sending state, just as E_r is of the receiving state. The distance between the two is captured by T_{sr}. But the price index term G_r also fits in nicely with the literature on gravity-models. It serves as a proxy for what has been called *remoteness* in Wolf (1997), the property that two regions will trade more than the simple gravity model predicts if the two of them are close, and relatively far from all other regions.[5] In this model, the two isolated states will have relatively high values of the price index G. This depresses their total trade, but relatively increases their bilateral trade. To see this, write the stream of goods in (5.10) as

$$\bar{D}_{rs} = n_s \frac{E_r}{G_r} \left(\frac{p_s^s T_{sr}}{G_r}\right)^{-\sigma}.$$

We see from the second factor that a high G_r causes the real trade spending E/G to be low. When this high G_r is caused by high values of T_{sr} as we assumed, this is not remedied by the fact that the price index enters with a positive power in the third term. However, if the value of the T_{sr} is especially low for a certain region s, trade between regions s and r is

---

[4] Due to the assumption of 'iceberg' transportation costs, to get the amount of goods in (5.9) to arrive in region r the producers in region s must actually ship T_{sr} times as much. This way, they account for the goods that 'melt' in transit. In many papers, this leads to an extra factor T_{sr} in this expression. However, iceberg transport costs are a convenient fiction and these extra goods are not observed in the data; it is therefore defendable to leave the extra T out, as we do in other chapters. Here, for consistency with other work, we maintain the extra T.

[5] Wolf uses as a measure for remoteness the ratio of bilateral distance to the average of the output-weighted mean distance to all other regions. This regressor is expected to have a negative sign.

relatively high. A specific case of two isolated states is examined below, on page 139. Related work that derives the notion of gravity and remoteness from a trade model based on CES-demand functions can be found in Anderson and van Wincoop (2003).[6] There, price indices G_i are referred to as "multilateral trade resistance," as they serve to measure the average impediments to trade for region i.

As we saw in formula (5.7) above, if we assume that each firm makes zero profits, each firm's production is fixed at certain level $\bar{Y}$. Because production, or supply, must be equal to demand, this introduces a relationship between the factors that influence total demand (formula 5.9) and the factors that determine a firm's price (formula 5.6):

$$\begin{aligned} \sum_{j=1}^{N} E_j G_j^{\sigma-1} T_{i,j}^{1-\sigma} &= \bar{Y}_i p_i^{\sigma} \\ &= \bar{Y}_i \left[\frac{\sigma}{\sigma-1} w_i^{\alpha} G_i^{1-\alpha} \right]^{\sigma} \qquad (5.11) \end{aligned}$$

This formula shows that there exists a negative relationship between a region's transport costs $T_{i,j}$ and its wage level w_i. Regions which are far away from large markets (and have a small market themselves) can be expected to have lower levels of wages compared to those close to the industrial core. This relationship forms the basis for several exercises, which study the model's relevance by looking at the correlation between a region's wage and its 'closeness' to other regions. As we saw in section 5.2.2 above, such a relationship is tested by Hanson (1998, 1999) for American counties and by Brakman et al. (2001) for German regions. Redding and Venables (2001) test the relationship using data on 101 countries worldwide, after they have approximated the term on the left hand side of this equation. We will take a closer look at this study below, where we use the same methodology on

---

[6] Anderson and van Wincoop (2003) do not explicitly model supply, but assume that each region produces one unique variety of goods. Demand takes a CES form over all goods and market equilibrium allows them to derive the gravity equation

$$X_{rs} = \frac{y_r y_s}{y_W} \left(\frac{T_{sr}}{G_s G_r} \right)^{1-\sigma}$$

where the notation is in terms of this chapter. Regional income y_s and y_r play the role of n_s and E_r in (5.10), normalized by world income y_W. The crucial difference between their model and the current model lies in the role of G_s, the price index in the sending state. In our model, a higher G_s implies higher costs of production, and thus a higher price: G_s appears in (5.10) mainly as a component of p_s^s, see (5.6). Thus, high values of G_s inhibit trade. In Anderson and van Wincoop, a higher G_s *encourages* trade as

> Higher barriers faced by an exporter will lower the demand for its goods and therefore its supply price p_i. (p. 176)

This result would be hard to obtain with a production model in which the price of intermediate inputs plays a role in the price of final production.

our own data. This will be the subject of in section 5.5.1. First, we look at the available data.

5.4 Data

The data used in this estimation concern 51 US states and are described in detail in the appendix on page 144. We make use of the Commodity Flow Survey 1997, a dataset compiled by the US Bureau of Transportation Statistics. It is an estimate of inter- and intrastate trade based on a survey among firms that ship traded goods within the US, using a variety of modes. Our other two main datasets concern wage levels in each state and the distances between states[7] and are discussed in the appendix. A fourth set, data on gross aggregate output per state, is used in this paragraph only. A map of the 48 continental states and their two-letter abbreviations is in figure 5.1 below.

Before estimating the model we specified above, we will use this data to estimate a simple gravity-equation. This simple model can serve as a benchmark with which we can compare the results. The model we estimate is

$$\log X_{rs} = \alpha_1 + \alpha_2 \log Y_r + \alpha_3 \log Y_s + \alpha_4 \log D_{rs} + A_{rs}\beta + \varepsilon_{rs}. \qquad (5.12)$$

This equation explains the flow of trade (X_{rs}) by total production in the receiving and in the sending state (Y_r and Y_s, respectively), the distance between the two states (D_{rs}) and a set of k dummy variables contained in the $1 \times k$ vector A_{rs}. We present estimates with different A's, where k varies between zero and two.

A possible dummy variable contained in A is the intra-state dummy, which is one if the sending and the receiving state are the same, and zero otherwise. Another possible dummy variable is the adjacency- or border-dummy, which is one if the sending and receiving state share a border, and zero if they do not.

The results are in Table 5.1. This simple gravity equation explains about 80 percent of the variation in trade flows. The elasticities to income α_2 and α_3 are almost equal and close to one. There is a substantial home bias in trade, measured by the intra-state dummy, and a positive effect of sharing a border. These two effects reinforce the negative coefficient on distance.

We compare our estimates with those in Wolf (1997) who uses an older and less complete version of the Commodity Flow Survey. Note that Wolf uses a form slightly different from (5.12), namely

$$\log X_{rs} = \alpha'_1 + \alpha'_2 \log(Y_{rs}^{\mathrm{comb}}) + \alpha'_4 \log D_{rs} + A_{rs}\beta' + \varepsilon'_{rs}. \qquad (5.13)$$

---

[7] Distances are computed as the length of a straight line between the weighted centers of two states, see appendix 5.A below.

Dep. Variable	Trade 97	Trade 97	Trade 97	Trade 93
Technique	OLS	OLS	OLS	OLS
Observations	2201	2201	2201	1030
Source	Own comp.	Own comp.	Own comp.	Wolf 1997
Constant	-8.843	-10.014	-11.354	-8.06
	$[-24.993]$	$[-29.262]$	$[-31.929]$	$[21.04]$
$\log(Y_r)$	0.992	0.998	1.004	0.961^*
	$[52.774]$	$[56.164]$	$[57.931]$	$[66.23]$
$\log(Y_s)$	0.995	1.003	1.011	
	$[53.377]$	$[56.925]$	$[58.804]$	
$\log(D_{rs})$	-1.171	-1.029	-0.864	-0.956
	$[-50.931]$	$[-43.890]$	$[-31.454]$	$[33.69]$
Intra state		2.076	2.492	1.338^{**}
		$[16.195]$	$[19.073]$	$[11.68]$
Adjacency			0.763	
			$[10.849]$	
R^2	0.787	0.810	0.820	0.867
SSR	1728	1544	1465	

* Wolf uses one regressor, $\log(Y_rY_s)$, instead of two. ** Note that Wolf divides the state-income variable Y by two if the trade flow is within the state itself (*i.e.*, the receiving and sending state are the same). This increases the intra-trade parameter. In our estimates, Y is not changed.

Table 5.1: Gravity equations

When the sending and receiving state are different, $Y_{rs}^{\text{comb}} = Y_rY_s$, but when the state trades within itself there holds

$$Y_{rs}^{\text{comb}} = \frac{1}{2}Y_r\frac{1}{2}Y_s$$

with $r = s$, so that the regressor in that case is equal to $2\log Y_s - \log(4)$.[8]

It seems that the parameter estimates between the two exercises are reasonably close. The only significant change can be seen in the parameter of the intra-state dummy. Our estimates show a higher value, and would have

---

[8] Without the exception for intra-state trade, the use of Y^{comb} amounts to the restriction that $\alpha_2 = \alpha_3$. Now, because the regressor Y^{comb} has a positive coeficient and is decreased for intra-state observations, the coefficient on the intra-state dummy is biased upward.

Figure 5.1: The 48 contiguous states in the US and their two-letter abbreviations.

been even higher had we followed Wolf's practice of dividing the GSP variable by two for intra-state flows. So where Wolf (1997) finds a home-bias effect, the later version of the same data shows an *increase* in that effect.

5.5 Estimation

5.5.1 The Redding-Venables approach

Redding and Venables (2001) use a closely related version of the above model to estimate its parameters in two steps. For this, they use data on 101 countries, including the trade flows and distances between them, whether they share a border and the level of wages in each country, approximated by GDP per capita. This data is used to estimate formula (5.10) with panel data methods as a gravity equation. Then, using the projected values for the terms in (5.10), they test the relationship between average trade costs and the level of wages in a region that we saw in formula (5.11). We will go over these steps in turn.

First stage estimation: gravity

For their initial estimation, the authors rewrite the trade relationship in formula (5.10) as[9]

$$X_{rs} = \phi_s T_{sr}^{1-\sigma} \psi_r \tag{5.14}$$

---

[9] As discussed in footnote 4, the authors add an extra T_{sr} to the equation. We will follow their convention here.

where ϕ_s is called country s's *supply capacity* and ψ_r is country r's *market capacity*. Each of these two terms contains information on a country's trade characteristics that is the same towards all its trading partners. Market capacity $\psi_r = E_r G_r^{\sigma-1}$ reflects the total amount of imported goods absorbed by country r. It increases when the country spends more on imports, or when it is (on average) far away from its trading partners.[10] Supply capacity $\phi_s = n_s p_s^{1-\sigma}$ varies with the number of firms in country s, and hence with its total production of tradeables.

Given the structure of formula (5.14), it is possible to estimate it using the fixed effects panel data method. We rewrite the equation as

$$\log(X_{rs}) = \delta_0 + \phi' \iota_s + \delta_1 \log(\text{dist}_{rs}) + \delta_2 \text{bord}_{rs} + \delta_3 \iota_s' \iota_r + \psi' \iota_r + u_{rs} \quad (5.15)$$

Once again, X_{rs} is the value of the flow of trade from region s to region r. The $N \times 1$ vector ι_i is filled with zeros, except at the ith position, where it is one. Thus, the $N \times 1$ vectors ϕ and ψ contain the supply and market capacities of all regions.

The dependency between distance and trade is captured by the δ-parameters. The first, δ_0, is a scaling factor. Distance (in miles) has a coefficient of δ_1, which we expect to be negative. The influence of the spatial characteristics of the two regions is further captured by two dummy-variables: bord_{rs} is one if the two regions r and s share a border[11] and the product $\iota_s' \iota_r$ is one only if the sending and receiving state are the same.

With data about the distances between regions, the size of their bilateral trade and whether they share a border, it is possible to estimate the parameters in relation (5.15). We shall do so for our data on the states in the US, described in section 5.4 above. For comparison, we also mention the outcomes of Redding and Venables (2001). They use data on 1994 bilateral trade flows between 101 countries. The distance between two countries is that between the capital cities. Trade within a country is not taken into account in their estimations, so the regressor $\iota_s' \iota_r$ is left out. In our dataset, data on trade within a state is available; we estimate both with and without it.

What can expect *ex ante* about the differences between the two estimations? Given that the methodology is exactly the same, variations in outcomes must be caused by differences between the two datasets. Firstly, the dataset of Redding and Venables is larger by a factor of four; *ceteris paribus*, this leads to smaller estimation errors. However, their data pertains to the whole world and is probably more heterogenous than that measured

---

[10] Countries with a small home market that are far away from trading partners will have a high value for G; this means that they will be less daunted by high import prices, since *all* of their import comes from far away.

[11] In the United States, some states share a border of size zero as their corners just touch each other. This is the case for Arizona and Colorado, for instance. In spite of this tangential relationship, the border-dummy is set to one for these pairs of states.

within the United States. For instance, the distance between two countries is likely to include a stretch of ocean, whereas this rarely happens between two US states. Given that trade over sea is more complicated, we expect higher trade costs in the World dataset. Also, two countries sharing a border is a more unlikely event than two states sharing one. This could make the effect of borders more significant in the World dataset. Finally, trade between countries may or may not be hampered by restrictions such as tariffs, or by cultural differences. Given the relative homogeneity of the US states, we expect less unexplained variation in the latter sample.

The results of the estimation are in table 5.2 on page 146. There are three estimations for both datasets, and two extra for the US dataset. We report the values for $\hat{\delta}_1$, $\hat{\delta}_2$ and $\hat{\delta}_3$, leaving the (large) vectors $\hat{\phi}$ and $\hat{\psi}$ out. These coefficients will be used later on, however.

In the first estimation (in the first two columns), the full sample is used. This includes pairs of regions for which no trade is recorded. For both datasets, this means that the actual trade between the two regions is probably very small. We substitute a zero for (the logarithm of) these unmeasured flows. We see that distance has the expected negative sign, whereas the border-dummy has a positive parameter. Both are highly significant. The coefficient of the variable $\iota_s' \iota_r$, called own_{rs} in the table, is also positive and significant. As expected, both distance and the occurrence of a border have a larger effect in the World dataset. The explained variance is about the same for both.

In the second estimation, pairs of regions between which no trade is recorded are taken out of the sample. This leads to smaller, but more significant coefficient estimates. For the World dataset, the R^2 does not increase; leaving out the zeros does not improve the performance of the model. The R^2 does increase, markedly, for the US dataset. This is caused by the fact that many unobserved pairs involve either Hawaii or Alaska, two states which turn out to be outliers in this dataset.

In the third estimation, we reintroduce the pairs with unobserved trade and treat them as left-censored observations. The model parameters are estimated using the Tobit method. This increases the coefficient on distance and decreases the border dummy. Standard errors are slightly worse, though.

The final two columns pertain only to the US dataset. In the fourth estimation, we use only contiguous states, eliminating Alaska and Hawaii from the sample. These two states suffer from many missing observations, whereas those that are available act as outliers. The District of Columbia is also struck from the sample, as the model also performs relatively badly for this region. This is probably due to its small size and atypical sectoral makeup. In the fifth estimation we eliminate the remaining 49 observations of in-state trade data. This hardly affects any parameters, showing that the use of an in-state dummy adequately captures the special nature of trade within the same state.

When we compare the parameter estimates for trade within the United States with those for world trade, at first glance the results rather similar. All corresponding parameters have the same sign and the order of magnitude is the same for similar parameters. The differences do amount to several times the standard error, though: the effect of distance and the effect of a shared border are greater for world trade data. The explanatory power of the model is greater for US data, however. Partly, this can be explained by the absence of administrative and physical barriers in the US. Also, the data on world wages is a proxy (GDP per capita), giving rise to extra measurement error.

Second stage estimation: Wages

We keep the results of the previous exercise to conduct a second stage estimation. For this, Redding and Venables construct two new variables: *Market Access* of a region s is defined as

$$\begin{aligned} \mathrm{MA}_s &= \sum_{r=1}^{N} E_r G_r^{\sigma-1} T_{r,s}^{1-\sigma} \\ &= \sum_{r=1}^{N} \phi_r T_{r,s}^{1-\sigma} \end{aligned} \tag{5.16}$$

and *Supplier Access* of region r as

$$\begin{aligned} \mathrm{SA}_r &= \sum_{s=1}^{N} n_s \left(p_s T_{r,s}\right)^{1-\sigma} \\ &= \sum_{s=1}^{N} \psi_r T_{r,s}^{1-\sigma}. \end{aligned} \tag{5.17}$$

The names of these variables suggest that they are not chosen at random. Market access is a weighted average of the expenditures on differentiated goods by the region's potential trading partners. The weights contain distance to the region with a negative sign and the relative isolation of the potential trading partner (as indexed by their price index G) with a positive sign. As such, the measure is reminiscent of the market potential function suggested by Harris (1954).

Supplier access in (5.17) is inversely proportional to the regional price index G_r, as defined in (5.4). It is an index of the ease with which firms in the region can get intermediate goods, and with which consumers can get final goods. The two variables defined above share two desirable traits: firstly, using the results from our first-stage estimation, we can compute their values. Secondly, they are related to the level of wages in a region and thus offer a way to test the model.

Computing the values of MA_s and SA_r involves using the estimated values of ϕ and ψ that we obtained earlier, and our estimate of the costs of transport. We construct

$$\begin{aligned}\widehat{\mathrm{MA}_r} &= \exp(\phi_r) \cdot \mathrm{dist}_{r,r}^{\delta_1} \cdot \exp(\delta_3) + \\ & \quad \sum_{s \neq r} \exp(\phi_s) \cdot \mathrm{dist}_{s,r}^{\delta_1} \cdot \exp(\mathrm{bord}_{s,r})^{\delta_2} \\ &\equiv \mathrm{DMA}_r + \mathrm{FMA}_r\end{aligned} \tag{5.18}$$

and

$$\begin{aligned}\widehat{\mathrm{SA}_r} &= \exp(\psi_r) \cdot \mathrm{dist}_{r,r}^{\delta_1} \cdot \exp(\delta_3) + \\ & \quad \sum_{s \neq r} \exp(\psi_s) \cdot \mathrm{dist}_{s,r}^{\delta_1} \cdot \exp(\mathrm{bord}_{s,r})^{\delta_2} \\ &\equiv \mathrm{DSA}_r + \mathrm{FSA}_r.\end{aligned} \tag{5.19}$$

Notice that we used our estimate of $T_{r,s}^{1-\sigma}$ from the previous section, which uses measures of distance, a border- and an own-state-dummy. In these formulas, we implicitly defined four other *access*-variables by splitting off access to the own region from access to other regions. DMA and DSA are domestic market- and supplier access, and FMA and FSA their foreign equivalents. Separating these terms will allow us to test them separately, later on.

To see how these measures of access interact with the wage level, write equation (5.11) as

$$\alpha\sigma \log(w_i) = \zeta + \log(\mathrm{MA}_i) + (1-\alpha)\frac{\sigma}{\sigma-1}\log(\mathrm{SA}_i) + \epsilon_i \tag{5.20}$$

for a region i. Notice that both market and supplier access have a positive coefficient in this equation. Products from a region with low market access incur large transport costs before they reach their customers. As these products have to compete with other, cheaper products, this limits the wages that can be paid in their production. Similarly, low supplier access means that intermediate goods are expensive: this squeezes the value that can be added in a region from the other side.

We will estimate equation (5.20) using generated values for MA and SA. These are computed as in (5.18) and (5.19), using predicted values for ϕ and ψ. This procedure renders OLS standard errors unusable: the stochastic errors in the gravity equation (5.15) turn up in the predicted values of MA and SA, which affect the stochastic behavior of ϵ_i in (5.20), violating the assumptions that underlie standard OLS analysis.

To estimate the standard error in spite of these difficulties, bootstrap methods are available (see Efron and Tibshirani 1993, for instance). For the gravity equation, we construct a new sample of the same size by drawing random observations (each observation a flow of trade and its regressors) from the original sample. This sample is a bootstrap-replication, for which original observations may be absent, or appear more than once. From the bootstrap-replication we re-estimate the trade-equation (5.15) and use the outcome to generate $\widehat{\mathrm{MA}}$ and $\widehat{\mathrm{SA}}$ as usual, which together with observations on wage make up a sample for equation (5.20). We generate 200 samples this way, the conventional number of bootstrap-replications according to Efron and Tibshirani. Of each of these samples, we use the same procedure to generate 200 bootstrap-replications. Estimating equation (5.20) on the resulting data gives forty thousand estimates, from which the standard error of the regressors can be directly observed.[12]

Several other problems potentially plague this estimation. As Redding and Venables remark, a contemporaneous shock to a region that affects both the independent variable and the regressors could introduce a bias the results. To eliminate the possibility of contemporaneous shocks, we estimate using wages from 1999 with regressors from 1997. This does not eliminate another class of 'third variables,' a time-invariant region-specific effect that plays in both a region's wage and in its market- and supplier access. To correct for this possibility, we report regressions on total access as well as 'foreign access,' as defined in (5.18) and (5.19). In the latter regressor, data from the own region does not play a role. Below, we will also add a number exogenous regressors that proxy for a region's time-invariant attractiveness and may capture its effect.

To start, we have to select a first stage estimation from the previous paragraph with which to work. We select the one that gives the best fit, called US 4 in table 5.2. This estimate uses the sample of all contiguous states, with trade flows including those to the sending state itself. As it turns out, the Market Access and Supplier Access regressors are highly colinear; the correlation between the two series is 0.95. This means that estimating (5.20) directly would be problematic. We proceed by using just Market Access as a regressor. At the end of the paragraph we compare the results to those obtained with Supplier Access.

The results of the estimation are in table 5.3 on page 147. We report the estimates on our US dataset, as well as the results obtained in Redding and Venables (2001, table 2). A scatterplot of the first two regressions for the United States is in figures 5.2 and 5.3 on page 148. Each point in the plot represents a state, indicated by its two-letter abbreviation. The horizontal

---

[12] As it turns out, bootstrap standard errors lie between one and two times the (invalid) OLS-standard errors, indicating that the extra variability due to generated regressors is reasonably small. We report only bootstrap-standard errors.

axis in figure 5.2 gives predicted market access according to formula (5.18). On the vertical axis the log of that state's average annual wage is plotted. Figure 5.3 is similar, only this time the variable on the horizontal axis is foreign market access.

From the first two columns, we note that the relation between foreign market access and the level of wages is much weaker in our estimation than in the *World* dataset. Both the explained variation and the statistical significance of the coefficient are smaller. The coefficient does have the right sign, however. From the scatterplot in figure 5.3 we can learn about the reasons for this weak performance. There is a clear positive relationship between FMA and wages for small states, such as Delaware (DE) and Vermont (VT). However, there are a number of outliers that spoil the correlation. These outliers consist of large states, whose own market is not a part of foreign market access. Especially those that are surrounded by (economically) smaller states fall outside the usual relationship, *e.g.* California (CA) and Texas (TX). This makes sense: explaining the wage levels in California by its proximity to Nevada and Arizona is bound to be problematic, but New Jersey's wage levels certainly have something to do with its wealthy neighbors.

The fact that relatively large states disturb our measurements may be an explanation for the fact that this estimation works better for worldwide data, where the dominance of large states is perhaps less of an issue.[13]

These problems disappear when we use full market access (MA) as a regressor, in the third and fourth column. The explained variance is about the same as in the World dataset, as is the statistical significance. This points to a large role for domestic market access, which is confirmed by the final estimation in columns five and six. Even though both coefficients have the correct sign, DMA clearly trumps FMA as a regressor for wages.

There may be a problem with the use of full market access as a regressor, though. As local demand in a state is included in this variable, local shocks that affect productivity in a state show up in the regressors as well as in the dependent variable. This causes simultaneity bias in the estimation.

Another detrimental effect of including local market access can be seen in the last two rows of table 5.3. There, we report the results of Moran's I test on the residuals of the estimated wage equation. Moran's statistic tests for spatial autocorrelation (see Cliff and Ord 1973, van Oort 2002, chapter 4) using a weight matrix to indicate which regions are close to each other. We use the matrix B as the weighing matrix, in which entries are equal to one if the two states share a border.[14] The diagonal of B consists of zeros. We

---

[13] According to the BLS (see appendix for data sources), at the end of 1997 California, Texas and New York together accounted for 25% of employment in the USA.

[14] The choice of the weight matrix is, to a degree, arbitrary and its impact should be measured. We have computed alternative statistics using a matrix B' where $b'_{ij} = \exp(-.001 \cdot \mathrm{dist}_{ij})$ (with dist_{ij} the distance between states i and j) and found that their level of signifi-

have used the data in B before, to estimate the trade equation (5.15).

Moran's I statistic is computed as

$$I = \frac{N}{\iota' B \iota} \frac{\epsilon' B \epsilon}{\epsilon' \epsilon} \tag{5.21}$$

with N the number of observations, ι a $N \times 1$ vector of ones and ϵ the $N \times 1$ vector of errors. In table 5.3 we also report the place of each Moran's I in the distribution of this statistic (under the hypothesis of no spatial autocorrelation).[15] All realizations of the statistic allow us to reject zero spatial autocorrelation at the 1,5% level, indicating that a high realization of the wage in one state makes a higher than expected wage in the bordering states more likely. However, the estimations which include local market access as a regressor show by far the most significant realizations of this statistic.

Are things any different when we use supplier access instead of market access as an explanatory variable? Our theoretical model tells us that SA and MA each determine part of the variation in wages, as can be seen in equation (5.20). However, we determined above that the pair of regressors suffers from severe multicollinearity and decided to include only measures of market access in the regression. By the same token, we could have decided to use only supplier access. The results of this estimation are in table 5.4.

Once again, we compare our results with those in Redding and Venables (2001). We see a similar pattern as in table 5.3: a regression using only foreign access gives a lower, and less significant, value of the coefficient and a lower R^2 compared to the World data set. Using a full measure of supplier access improves the estimation but leads to higher spatial autocorrelation in the residuals.

We will try to improve these estimations below by adding data on the exogenous amenities to productivity that characterize each state, as well as by employing instrumental variables in our estimation.

Exogenous amenities

When we estimate state-level wages as a function of market- and supplier access, we neglect all other factors that may also have a bearing on those wages. In as much as these factors correlate with our regressors, they can

---

cance was very close to the values obtained with B.

[15] The expectation of Morans I is $-1/(N-1)$, with N the number of observations. We bootstrap the distribution of I by generating 100,000 vectors ϵ^*, where each ϵ^* is a random permutation of ϵ (in the usual terminology of spatial autocorrelation, we use *nonfree sampling*). We compute the corresponding values of I, and indicate the percentage of outcomes *higher* than the recorded statistic. An asymptotic distribution for the statistic is known (Cliff and Ord 1973, chapter 2) but its small-sample behavior inspires more confidence in bootstrap methods (see Anselin and Florax 1995).

cause a bias in the estimation. It is easy to think of a situation in which this may happen.

In the first paragraph of this chapter we mentioned that so-called 'first-nature' causes of geographic concentration also play a role: the physical features of the area, its climate and natural infrastructure may have an effect on productivity. Imagine, for instance, that a predominantly warm climate opens up economic opportunities (*e.g.*, tourism) in a state. This may raise the general level of wages. If a number of neighboring states share the same climate, this 'third factor' will increase wages in all of them. Being close together, market- and supplier access for each of these states will probably be at a comparable level. Suppose it is lower than average; in that case, the unobserved regressor 'climate' causes a downward bias in our estimates.

In order to test the robustness of our initial estimates against the influence of third factors, this section presents the results of a number of regressions similar to those above, but including a number of possible third factors as regressors. We use the following exogenous amenities:

- **Climate**. In order to control for an exceptionally warm or cold climate we use two regressors, normal yearly heating degree days (*nrmhdd*) and normal yearly cooling degree days (*nrmcdd*). The former is defined as the cumulative number of (Fahrenheit) degrees in a year by which the mean temperature of each day falls below 65°F, the latter as the cumulative number of degrees in a year by which the mean temperature lies above 65°F.[16] The idea is that an exceptionally warm or cold climate may account for differences in productivity. For reasons of scale, we divide these regressors by 1000 in the actual regression.

- **Geology**. Special economic opportunities may arise from the presence of precious minerals in a state. To proxy for these opportunities, we use the value of nonfuel mineral production per square kilometer in 1997, as reported by Smith (1997), in thousands of dollars.

- **Access to sea**. Finally, we include a dummy variable that indicates if there exists a deep sea port in the state. Access to sea may proxy for the possibility of international trade.

We expect heating- and cooling degree days, regressors that indicate an unpleasant climate, to have a negative impact on productivity. The presence of minerals is likely good for wages, as is the presence of a port. We first regress wages on these exogenous amenities alone, and then include our measures of market access. The results are in table 5.5.

---

[16] Somewhat counterintuitively, cooling degree days measure warmth and heating degree days measure coldness. An example may clarify: if the mean temperature in a state is 67°F all year long, the yearly cooling degree days are $(67 - 65) \times 365 = 730$ and the yearly heating degree days are zero.

From the first column of this table, we notice that the four exogenous regressors have the expected sign and succeed in explaining about half the variation in wages. However, Moran's I is rather high (higher than all but 2.8% of the distribution under H_0) and indicates that we may not have all region-specific exogenous amenities in our set of regressors. The inclusion of Foreign Market Access in the regression hardly changes the values of the earlier coefficients. However, the value of the coefficient for FMA is about a third of the earlier measure (table 5.3) and lies below one standard error. Explained variation hardly improves. Including FMA does improve Moran's statistic to a point where we are no longer able to reject the hypothesis of no spatial autocorrelation at the 5% level.

Things turn out differently when we include measures of (Domestic) Market Access. The coefficients of the exogenous regressors change substantially (more than one standard error in all cases) and Moran's statistic again increases to a significant level. This result again points to problems with the inclusion of Domestic Market access.

IV estimation

The estimations above may suffer from the occurrence of simultaneity bias, which occurs when the error term from an estimation is correlated with one (or more) regressors. In this matter our model is clearly the culprit, as it indeed allows the error terms to influence the market access regressors. We discuss how this happens and how we can correct for it. After that we assess the size of the problem.

The error terms in the regression imply that observed wages are, to a degree, inconsistent with our model, either because of measurement error or because of misspecification. Where we expect a wage w_i^* in state i, we actually find $w_i = w_i^* + \epsilon_i$. That w_i is the dependent variable in our estimation, but it also makes its way into the regressors; according to (5.6), prices are a function of the wage and via (5.8), those prices end up as an element of all the price indices $G_r, r = 1, \ldots, N$. Our regressors, MA and SA, are again a function of prices and price indices (*cf.* formulas 5.16 and 5.17). This puts the error ϵ_i in the (supposedly) exogenous variables. The question is, whether the weight that ϵ_i receives in MA_i and SA_i is large enough to influence the estimation.

With this problem in mind we used two regressors above, MA and FMA, where the former excludes market data from the own state. The use of local market capacity in the regressor MA will probably introduce ϵ in MA with a large weight. Indeed, we find that the regressions where FMA is used instead of MA show lower spatial autocorrelation of the errors.

However, we can also eliminate ϵ from the regressors entirely if we employ instrumental variable estimation. This idea is used in Mion (2003), who takes a panel-approach on Italian data. Brakman et al. (2004) use it to isolate the effect of one particular disturbance in a spatial growth process and Ciccone and Hall (1996) employ four "deep historical" instruments that proxy for the innate attractiveness of American states as places of residence.

For IV, we need instruments that correlate with the regressors MA and SA, but not with the errors ϵ. Once again following Redding and Venables (2001) we use distance from major economic centers as instruments, in particular the distance from New York City and from Los Angeles.[17]

The results are in table 5.6. The first two columns use only Market Access variables as regressors, and can be compared to the results in table 5.3. Note that the coefficient of FMA is comparable, while MA has a coefficient that is much lower than before. It appears that simultaneity bias plays an important role in the estimations which use this regressor.

The last two columns once again make use of the exogenous amenities that were introduced above, and can be compared to table 5.5. Here also, the coefficient for MA has fallen. Note that we can no longer reject the hypothesis of no spatial autocorrelation in the third column. On the other hand, the significance of Market Access as a regressor is tenuous.

Discussion

We have estimated a relationship that explains the levels of wages in the United States by the level of market access. The variable that indicates market access is itself a construct from the results of a regression, which resembles a gravity-type relationship. To construct the measures of access, heavy use was made of the theoretical model of economic geography.

The estimations mimic those of Redding and Venables (2001), but the results are less satisfying. To a certain extent, this can be explained by the nature of our dataset: it is smaller and possibly more dominated by large regions. However, the fact that we use data on US states also brings some advantages, which fail to realize. For our regressor, we are able to use actual recorded wages instead of a proxy.[18] Also, institutions are bound to be more similar inside the USA than worldwide. This means that institutional differences (and, for that matter, international frictions such as tariffs) are no longer a factor. These differences were proxied for by distance, but supposedly less than perfectly. In spite of these advantages, the explanatory power of our model, especially when it relies on foreign market access, is less than that measured on a worldwide scale. The same hold when supplier access is used as a regressor.

---

[17] As usual, distance is measure from the (employment-weighed) center of the state so that New York and California each have positive distances to these economic centers.

[18] Redding and Venables (2001) use GDP per capita for their main estimations, altough they do estimate the relation with wage data for a smaller sample.

Our initial estimations suffer from an omitted variable bias that results in spatial autocorrelation of the errors. We remedy this problem by introducing an extra set of regressors that proxy for exogenous qualities of each state, such as climate and infrastructure. Furthermore, we estimate using distance to economic centers as an instrument. This ensures that the active element in the Market Access variables is indeed the access to markets in other states. These estimations show that the explanatory power of the model is present, but limited.

A potential problem with the methodology used above is the fact that the estimated relationships are not neccesarily consistent with the general equilibrium solution of the model. For instance: when we start off estimating the gravity equation in (5.14), we parametrize the relation in (5.10). The latter shows that each region's supply capacity is directly related to the number of firms and the price, both of which are in turn determined other variables in the model, as seen in (5.6) and (5.7). The same goes for market capacity. However, this relationship is not used in the procedure until much later: only when we regress regional wages on the access variables do we observe that in fact, the relationships of the model do not hold: if they did, the regression would have had to give us a perfect fit. The variables that were kept constant would not, had they been subjected to the rules of the model, have stayed so.

This leaves us with the question of how to interpret the findings in this section. On the one hand, we have shown that there exists a significant correlation between regional wages and access variables. This is an indication that the model has some explanatory power in our dataset. On the other hand, the less-than-perfect correlation between wages and access variables shows us that some of the relations inside the model are violated; with our current methodology, we have worked from one end of the model (the trade relationship) towards the other end, leaving all discrepancies to accumulate along the way.

There are many plausible reasons why, even if the real world were governed by this model, we could not hope for a perfect correlation in our final regression. Measurement error, for instance, or the imperfect approximation that we use for transport costs. It remains slightly unsatisfying, however, that the numbers that we use for our estimation are not necessarily an equilibrium outcome of the model. This is especially true in the class of economic geography models, where for certain parameters a distributed outcome is infeasible, and agglomeration the only stable solution.

This is why we estimate the same model in a different way in the next section. The procedure there takes the general equilibrium nature of the model seriously and allows all the relationships to hold. We will compare the outcomes of the two procedures and use them to judge the importance of keeping variables constant.

5.5.2 General equilibrium estimation

In most estimations of the economic geography model, part of the model is kept constant. This leads to outcomes that are interesting, but not model-consistent. For instance, in our estimation in the previous section, the variable n and by extension E are kept constant, even though they are functions of other variables in the model. In this section we will attempt to estimate parameters while maintaining general equilibrium.

The properties of our model are governed by a small number of key parameters, the most important of which are transport costs and the elasticity of substitution. As we saw in paragraph 5.2.3 above, it is not always possible to separately identify these two parameters from the data, if flows of trade are the only information. If both parameters affect these flows in a similar way, it is impossible to separate out their influence. However, when data on wages is also available, this difficulty can be overcome. We will show this in the following paragraph. After that, we will estimate the key parameters in our model.

Identification

Bröcker (1999) is not able to separately identify transport costs parameter τ and elasticity of substitution σ. This is intuitive: a higher elasticity of substitution means that goods have become less differentiated. Given that products from another region are relatively expensive because of transport costs, if they become less 'special', their consumption will decrease. Thus an increase in either transport costs or the elasticity of substitution will have the same effect, hence it is not possible to identify the parameters separately.

Consider the model of the previous section and assume for a moment that there are only two regions. The value of the flow of trade from region 2 to region 1 will then equal

$$D_{21} = \frac{n_2 p_2^{1-\sigma} (T_{12})^{1-\sigma} E_1}{n_1 p_1^{1-\sigma} + n_2 p_2^{1-\sigma} (T_{12})^{1-\sigma}} \tag{5.22}$$

(this is a version of equation 5.10). Now assume that $\mu = 1$, so there are only tradable goods, and $\alpha = 1$ so that the production process only uses labor as an input. We can then write $n_r = L_r/\sigma F$ and $p_r = w_r \cdot \sigma/(\sigma - 1)$. Equation (5.22) reduces to

$$D_{21} = \frac{w_1 L_1}{\frac{L_1}{L_2} \left(\frac{w_1}{w_2}\right)^{1-\sigma} (T_{12})^{\sigma-1} + 1} \tag{5.23}$$

The above equation shows that if there is no data available about the level of wages in the different regions, w_r, then separate identification of T_{rs} and

σ is impossible, as the variables only show up in a joint term. The analysis by Bröcker (1999) was carried out without data on wages, so that implicitly all wages were assumed equal. This explains why there were identification problems in that study.

If we revoke some of our assumptions, identifiability is less clear. For instance, if we allow the share of intermediate products $1-\alpha$ to be positive, do we still need variation in wages for separate identification of the parameters? We can no longer tell from the simple expression for D_{21}, as we need to include the more complicated terms buried in the p's and n's of (5.22).

To find out about the possible separate identification of σ and τ we run the following simulation experiment: for a model with three regions, we generate 2000 random distances and labor supplies. This sample is split in two halves. For the first half, we generate random (different) wage levels for the three regions; for the second half, all wages are assumed equal to one.

For all 2000 sets of data we compute the equilibrium flows of trade as described above.[19] Because there are three regions, this is matrix of nine flows (both intra- and inter-region). We rearrange these flows into a vector T.

It is possible to compute the derivative of this vector T to changes in σ and τ. We do this numerically, so that

$$T_\sigma \doteq \frac{T(\sigma + \mathrm{d}\sigma) - T(\sigma)}{\mathrm{d}\sigma}$$

where $\mathrm{d}\sigma$ is a small number, on the order of $0.01 \cdot \sigma$. The operation for τ is similar. These two derivatives (two elements of $\mathcal{R}^9$) are compared by computing the angle ϕ between them. From linear algebra, there holds that

$$\cos(\phi) = \frac{< T_\sigma, T_\tau >}{\sqrt{< T_\sigma, T_\sigma >}\sqrt{< T_\tau, T_\tau >}}$$

with $< ., . >$ the inner product of two vectors. An angle close to zero means that there exists colinearity between the two derivatives, which could be a sign of identification trouble. Ideally, the angle between the two derivatives should be $90°$, indicating that the two regressors are orthogonal and perfectly identified.

Of course the derivative is only a first-order approximation to the nonlinear problem that we are trying to solve. However, the numerical derivative that we examine in this experiment will likely also be used by the solver-routine that computes the estimation. If we find closely correlated

---

[19] We use $\sigma = 6$, $\tau = 0.08$, $F = 1$, and $\alpha = 0.6$. These are fairly typical values. The results of the experiment are robust to variations in this set. When Z is a random variable with a standard normal distribution, we draw $w = 6 + 2X$, $Ls = 10 + 2X$. Each region is given coordinates (x, y) which both are draws from X. Distances are then computed by Pythagoras' theorem.

regressors in this case, it will certainly make estimation very difficult and lead to imprecise estimates.

The results of our simulation experiment are in figure 5.4. The two panels of the figure are histograms of the distribution of the angle between the two regressors. If we take as a rule of thumb that identification is feasible if the angle is larger than 20°, we see that most cases in the bottom panel (where wages vary) pose no problem. In the top panel, we see that identification of both σ and τ is a more rare event. In either case, identification is not clear cut—a bad realization of the regressors can throw a wrench in the works at any time.

Equilibrium Estimation

In this section, we will attempt to estimate the parameters of the model using a method in which full equilibrium is maintained. For this estimation, we use largely the same dataset as in section 5.5.1. The sources of this data are described in an appendix on page 144. As before, we take the 48 contiguous US states as our sample; of these states, we take as exogenous the matrix of distances between them, and a matrix of dummies indicating whether a border exists between two states. Our sample year is 1997: we use the amount of employment in each state and average state wages for that year as exogenous inputs.

The model is described in section 5.3. Given the exogenous variables above and values for its parameters, we can numerically compute a solution to the model in which equilibrium is attained on each market. That is, firms make zero profit, demand and supply are equalized for each product and prizes reflect a marked up average of the costs of labor and intermediate products. From this equilibrium we can compute a flow of trade between every pair of regions.

Our estimation procedure finds the parameters that make these flows of trade as close as possible to the observed data. It works as follows: we compute the outcome of the model as described above, given a set of values for the model's parameters. From this outcome, we take the matrix of trade flows and compare it to the matrix of actual flows of trade, available from our dataset. We then evaluate how close the two are (a discussion of how we measure closeness is below) and repeat the procedure for another set of parameters. This way, we search for the parameters that generate a flow of trade that is closest to the actual data. Though we are not able to exactly replicate the actual flow of trade in our model, we do find the model that comes closest while still being consistent.

There are a number of problems with the above procedure. First of all, our matrix of trade flows is incomplete: some of the state-pairs have an unknown flow of trade between them. We leave these empty observations out of the sample.

Secondly, the number of parameters that has to be specified is reasonable but too large to estimate all using this procedure.[20] In order to compute the equilibrium, we need to specify values for the elasticity of substitution σ, labor share α, industrial consumption share μ, firm fixed costs F and the two transport costs parameters τ, and the coefficient for the border dummy, δ. Transport costs are computed as

$$T_{i,j} = \exp(\tau \mathrm{dist}_{i,j} + \delta \mathrm{border}_{i,j}). \tag{5.24}$$

Our estimation consists of the numerical minimization of the objective function over all possible parameters. In order to cut down on computer time, we will formulate the model so that only three parameters have to be estimated.

First of all, we recognize that the fixed cost parameter F does not influence the outcome. The amount of fixed costs that is involved with starting a firm will influence the number of firms n linearly, as can be seen from (5.7). However, since fixed costs are the same in every region, the relative number of firms between regions stays the same and the pattern of trade is not affected. We set $F = 0.01$, a value that leads to numerically efficient values of n.

Next, we specify our objective function so that the absolute value of the flows of trade is not relevant. This will allow us to keep the share parameter μ out of the estimation. Finally, we get the share of labor in the production function from actual observation rather than this estimation. Table 5.7 gives the factor shares in production from 1997. We use our assumption that only labor and intermediate goods are used and compute labor's relative share from that, and set $\alpha = 0.605$.

We are left with three parameters to estimate: σ, τ and δ. The object of the estimation is to find the values for these parameters that give a model solution in which the simulated flow of trade is as close as possible to the observed flow of trade. We operationalize this criterion in the following way, dealing at once with missing observations and matters of scale: we start with the matrix of trade flows that is generated by the model, and which has the same dimensions as the flow of trade-matrix from the data (48×48). From this matrix, we set all entries for which the corresponding entry in the data-matrix is missing, to zero. In the data-matrix, these missing entries are also set to zero. We then erase all the diagonal entries, which are the within-state shipments. In the previous section, we found that these flows of trade behave differently from interstate flows, so that the dummy variable own was necessary. Leaving these observations out frees us from having to measure a parameter for the effect of the own-dummy, leaving us with only three parameters to estimate.

---

[20] While making point-estimates would be possible, the Monte Carlo error analysis below would become rather involved.

Next, both matrices are normalized so that the row-totals are equal to one. Because the matrices are drawn up so that each row contains all the imports of a particular region, this converts the entries to *shares of observed import*. That is, we express each flow of trade as a fraction of the total measured imports of the receiving region. Both matrices have a number of zeroes in them, which indicate that some entries were erased or not observed; as we shall see, these zeroes do not hinder the estimation process.

We can now formulate the criterion for closeness between the two matrices. Our estimation will minimize the sum-of-squares criterion

$$C(\sigma, \tau, \delta) = \sum_{i=1\ldots48} \sum_{j=1\ldots48} \left(D_{i,j} - P_{i,j}(\sigma, \tau, \delta)\right)^2, \tag{5.25}$$

where D is the data matrix of observed import shares and P the model's projection of the same matrix. Notice that missing entries in D and P, which have both been set to zero, do not add anything to C.

By minimizing the difference between the two matrices of shares, we are able to leave the consumption share parameter μ unspecified; its values does not influence the outcome of P as it affects both the single flows of trade and total imports in the same (multiplicative) way. Also, we bring balance between the data on imports from large states and data on small states; by using shares instead of dollar values, large states do not dominate the estimation.

The results of the estimation are in table 5.8. The values for the parameters in this table minimize, in terms of the criterion in (5.25), the difference between the projected import share matrix and the observed version. We have used an adapted Monte Carlo method to arrive at an estimate for the standard error of the coefficients. For this method, we use the projected import share matrix $P(\sigma^*, \tau^*, \delta^*)$ and the set of projection errors $[D - P(\sigma^*, \tau^*, \delta^*)]_{i,j}$ for observed pairs (i, j).

In a normal Monte Carlo procedure, we would randomly draw errors from this set and add them to P to generate a new dependent variable. Using this new dependent variable, we would re-estimate the parameters in a replication of the original estimation procedure. The variation in a set of about 200 replications of parameters re-estimated in this way would be an indication of the standard error of the original estimate.

However, in this case we run into trouble using the above procedure: it is quite possible that some of the newly generated dependent variables contain a number of negative entries, where a large negative error was added to a small initial projection. Beside the fact that negative trade would never be observed in practice, our model can not hope to replicate this observation as its trade is positive by definition. One could choose to leave the negative observations out of the replicated sample, but this leads to another problem: because small observations have a much higher probability of being left out, the new estimation is biased. This is observed in the fact that the average of the Monte Carlo replications is quite different from the original parameter estimate.

We therefore modify our Monte Carlo procedure as follows: firstly, we divide the set of projection errors by 40 before embarking on the original procedure. These smaller errors do not lead to negative dependent variables anymore. We observe the variation in our replicated parameter estimates in the usual way and then rescale it to take account of our initial attenuation of the errors. This is done by multiplying the observed standard errors by 40. It is these standard errors that are reported in table 5.8. We also plot the distribution of the Monte Carlo parameter estimates *before* correcting for the attenuated errors; these plots are in figures 5.5 through 5.7. From these figures, we notice that even though the estimation procedure is nonlinear, the distribution of the errors appears close to normal.

Our dependent variable in this estimation has been observed import shares. It is a natural variable to choose, as our model delivers these shares from the simple formula

$$\phi_{rs} = n_s (T_{rs} p_s^s)^{1-\sigma} G_r^{\sigma-1} \tag{5.26}$$

with ϕ_{rs} the share of region s goods in region r's imports. In principle, however, we could also have looked at *export* shares. These shares, ψ_{rs}, follow from our model as well:

$$\psi_{rs} = \frac{E_r n_s (T_{rs} p_s^s)^{1-\sigma} G_r^{\sigma-1}}{\sum_{r=1}^{N} \left(E_r n_s (T_{rs} p_s^s)^{1-\sigma} G_r^{\sigma-1} \right)}$$

but the expression is much more involved. Estimation on the basis of ψ_{rs} can be carried out through simple rescaling, however, if we use observed total exports and total imports for each region and use the restriction that these totals must hold for the projected trade matrix as well.[21]

The results of this second estimation are in the second column of table 5.8. Notice that while the parameter estimates appear to be robust to this transformation, the standard errors have increased. This can be explained as follows: in the first estimation we only use the model to derive import shares, which are a simple function of prices and transport costs (see formula 5.26). There, we do not use the model's information on the relative size of regions. In the second estimation, we have made an added assumption that involves the (economic) size of regions; this assumption is embodied in our transformation of import shares into export shares, and the added assumption has decreased the model's fit on the data. For our

---

[21] This way, we forego the use of expenditure variables E_r from our model. In a projection, the import shares are computed as in (5.26), and rows are rescaled to sum to total observed imports. Then, columns are rescaled to sum to one. This matrix is compared to the actual matrix of export shares.

analysis of counterfactuals below, we therefore use the parameters from the first estimation (import shares) only.

The table with results also reports the R^2 of the estimation. This number is defined, as usual, to be a measure of explained variance. We compute the total sum of squares TSS as sum of squares of $D - \bar{D}$. The matrix $\bar{D}$ is constructed as a matrix of ones and zeros, where the entries indicate if an observation is present at that position in D. Each entry is then divided by the sum of its column. The matrix $\bar{D}$ is our best guess for D if no data is used; it is a generalized intercept of the estimation. We also compute the residual sum of squares, RSS, as the total sum of squared errors $D - P$. Then, $R^2 = 1 - RSS/TSS$.

We would like to be able to compare the results of this estimation to the results of section 5.5.1. Both approaches try to fit a model in which the flow of trade is explained by costs of transport and the economic size of regions, a relationship which we wrote down in formula (5.10). The two approaches differ, however, on their specification of the transport cost function. The estimation in the previous section used formula (5.15), where the log of trade is proportional to the log of distance. Therefore, the parameters in table 5.2 give $\partial \log(X_{ij})/\partial \log(\mathrm{dist}_{ij})$ and $\partial \log(X_{ij})/\partial \mathrm{border}_{ij}$. In this section, we specified the transport cost function (5.24), which is used in formula (5.10). Combining these two tells us that the parameters in table 5.8 give $\partial \log(X_{ij})/((1-\sigma)\partial \mathrm{dist}_{ij})$ and $\partial \log(X_{ij})/((1-\sigma)\partial \mathrm{border}_{ij})$. In table 5.8, we therefore include comparable parameter estimates. For these estimates, the distance coefficient is $\hat{\tau}$ multiplied by $(1 - \hat{\sigma})$ and by the average distance used, which is 997 miles.[22] The border coefficient is $\hat{\delta}$ multiplied by $(1 - \hat{\sigma})$.

When we look at these comparable parameters, we should compare them to the outcomes of column US4 in table 5.2, which was estimated on the same sample. We see that when import shares are used, trade is more responsive to distance in the equilibrium estimation. When export shares are used, trade is less responsive to distance. The border parameter is greater in both equilibrium estimations.

Remoteness

With the parameter estimates in, we can use our model to take another look at the remoteness-effect that was described on page 117. The effect tells us that two regions trade more than a gravity-model would predict if they are relatively close together, and relatively far from the other regions. This follows from the fact that the gravity model fails to take into account a region's

---

[22] The latter multiplication is necessary because the use of distance in one, and log of distance in the other regression. It follows from the rule that $\partial \log(x) = \partial x/x$, but is only an approximation.

other possibilities for trade, when looking at one particular trade relationship. For instance, the prediction of the gravity model for trade between the Netherlands and Belgium is the same, whether these two countries are part of continental Europe or whether they are twin islands in the middle of the Pacific. Our model does account for other possibilities through the price index of manufactured goods G, which is high for remote regions.

We use our estimated parameter values (import shares) to compute the price indices G for the states of Oregon and Washington. They are 40% and 51% above the national average, respectively, indicating that these two neighboring states both occupy remote positions in the USA. If only distance and size matter, we can predict the trade from Washington to Oregon with the simple gravity equation (5.12). Using our estimation results from table 5.1, we find that predicted trade would be \$ 2125 million. The Venables-model predicts a much higher trade, as expected: the prediction is \$ 6392 million. This is due to the *remoteness*-effect that was discussed above. Both models grossly underestimate the actual trade, incidentally, which is \$ 10301 million.

5.6 Evaluating counterfactuals

In this section, we use the parameter estimates of paragraph 5.5.2 above in our model to compute the effect of two changes in the economy. First, we look at the spatial impact of a fall in wages in Illinois, a simulation exercise that is similar to an experiment conducted by Hanson (1999). We then use our model to compute the effect of a new interstate highway that causes a fall in transport costs.

5.6.1 A wage shock

In Hanson (1999), the author estimates the coefficients of a model where the level of wages in a region depends on economic activity in the surrounding regions, with distance reducing the influence. Hanson then studies the effects of a reduction in personal income in the state of Illinois, one of the regressors, on wages around the country.

In this section, we conduct an experiment that is similar in spirit. Since wages in our model are fixed and personal income is computed endogenously, we reverse the shock: wages in the state of Illinois are decreased by ten percent, and we observe the effects of this change on demand in every state. We pick Illinois for the same reasons as Hanson: its large economic size, which gives us a sizeable effect, and its central location.[23] Also, we can compare the range of this shock to that of the shock in income.

---

[23] In the map in figure 5.1 on page 121, Illinois is indicated with IL.

How does a reduction in wages in one state affect the demand for industrial goods locally, and in other states? In the afflicted state, prices drop as they are a markup over, amongst other, wages (formula 5.6). The cheaper goods from this state reduce the price index of industrial goods in every other state, but more so if the other state is close to Illinois. This causes the other states to lower their prices as well (formula 5.6 again) which sets off another round of falling price indices around the country. This process continues until convergence.

In the state where wages are lowered, two other effects come into play. The number of firms in the state decreases (see formula 5.7, where w goes down more than G). And, with state spending in this model directly related to the level of wages, the demand that the state itself exerts drops with 10%. Both effects depress the demand for goods from this state, but they are offset by an increase in demand from other states. This increase is the result of the lower price of industrial goods.

The change in prices and price indices leads to a change in the demand for each state. The effect is greatest for the state in which wages went down, because the price change is greatest in this region. Other states face two opposite effects: because all prices are down, a wealth effect causes an increase in demand for all states. However, their prices have all increased relative to Illinois which deflects demand away from the other states. We noted that states close to Illinois saw the greatest fall in prices, but they also suffer the most from Illinois' lower import demand.

Table 5.9 summarizes the results of this simulation. We look at the effects on several variables in Illinois itself, the neighboring states, a group of states at 'average' distance and two faraway states.

The total demand for industrial goods is up in Illinois, and down everywhere else. For states close to Illinois, the fall in that state's imports plays a major role in the drop in demand. For states further away, the deflected demand due to lower Illinois prices is the main cause. This can be seen from the second column in table 5.9, which gives the change in demand when we keep the imports in Illinois constant, and just look at the effects of the changed prices. This makes a big difference for states close to Illinois, who were able to lower their prices due to the cheaper inputs from their neighbor.

The number of firms falls in Illinois, which is a result of our zero-profit condition (5.7). Everywhere else, this number increases due to the fall in costs of input. Once again, the neighbors see the biggest increase.

The price index of industrial goods is down in every state. In Illinois itself, the change is relatively small due to the lower number of firms in that state (leaving it with less varieties). For other states, the fall in G with constant wages leads to a fall in prices, as can be seen in column 6.

Our conclusion that only the state with lower wages gains in demand obscures the fact that some states do see an increase in demand from certain trading partners. For instance, the final column of table 5.9 shows the change in demand from Florida, a state that is relatively far away from the Midwest. We see that Florida increases its imports from (cheaper) Illinois sharply after the shock, offsetting the change with a decrease in imports from other states. However, it also starts importing more from the states *around* Illinois, who are also able to lower their prices. So, these neighboring states see an increase in demand from trading partners that are far away, but the net effect remains negative.

What is remarkable about the effects of the shock, is that they do not seem to decline with distance much. States that are far away see a change in demand that is not smaller than the change experienced by intermediate states. Neighbors have a slightly different experience, but in the end see a similar drop in total export demand. This stands in marked contrast with the results of Hanson (1999), who measured that the effects of a ten percent drop in expenditures in Illinois reached no further than 900 kilometers. We must note that a drop in expenditures in his model leaves prices unchanged. The change in prices was the driving force behind the effects in our model.

We also note that the stability of the model owes a great deal to the assumption that the number of workers in each state remains constant. Presumably, with the changes in wages and the entrance and exit of firms, people would enter or exit the labor market. We have abstracted from this effect, possibly underestimating the results of the shock.

5.6.2 A fall in transport costs

We now experiment with a fall in the costs of transport between two states. Once again our change takes place in the Midwest. Assume the construction of a new highway between Illinois and Indiana, two neighboring states, that reduces the costs of transport between the two states by fifty percent.[24] For each state, it also reduces the internal costs of transport by half. We proxy for costs of transport by distance, so in effect we reduce the three relevant distances in our model.

Lowering the costs of transport between the two states affects either's price index of industrial goods: for firms and consumers in Indiana, products from their own state and those from Illinois become less expensive. Firms and consumers in Illinois see their own products and those from Indiana become cheaper. This means that the price of inputs drops in both states, allowing firms to lower their prices and new firms to enter.

We do not model the effect that the new highway has on traffic passing *through* the two states, on its way from Ohio to Missouri, for instance, or the effect on traffic reaching either state from the outside. This means that

---

[24] In the map in figure 5.1, Indiana is IN.

the only thing that the other states notice from the new road is the change in Illinois' and Indiana's prices. This change in turn lowers their average price for inputs and their prices for the final good, setting off another round of price drops. In the end, the new prices cause a shift in demand.

How exactly this shift plays out for each state can be read from table 5.10. We see that the two states which received the shock see an increase in demand. Their neighboring states are able to lower their prices, which would have earned them additional business, if it had not been for the competition from Illinois and Indiana. We see that the neighbors see little change in the demand from faraway Florida (column 6), but suffer an overall decline in demand. This decline is the result of Indiana and Illinois spending turning inward and to each other, away from the neighbors.

Intermediate states notice that their competitiveness *vis à vis* the Midwest has gone down, which causes demand to fall. The same holds for states that are far away, although we see that this time, the effect does become smaller with distance.

We note that the way we have modelled the effects of a new road is rather coarse: only the two affected states notice the change in their costs of transport. A more realistic modelling of the change would have taken the effects on other entries in the distance matrix into account. Such a computation is beyond the scope of this chapter, however, as it requires a complex model of transportation. In this section, we merely aim to illustrate the effects of a change in transport costs, and the mechanisms involved. In chapter 6 below, we do employ the results of a complete model of transport costs when evaluating the effects of a new railway link in the Netherlands.

5.7 Conclusions

In this chapter, we have estimated the parameters of an economic geography model in two ways: one method, previously employed by Redding and Venables (2001), assumes certain parts of the model constant and measures the correlation between market access and wage. A second method finds the model's general equilibrium solution that best reproduces the data. Both methods have been applied to a dataset that covers the US states in 1997. The dataset is described below, in appendix 5.A.

We find that the model gives a reasonable description of the trade between US states. The 'foreign' regressors in the first method performed a bit worse than they did on a worldwide dataset. The general equilibrium estimates gave reasonable values for the model's parameters, but their standard errors leave room for doubt.

When we simulate counterfactuals with out calibrated model, we see that the effects of a shock reverberate throughout the US. The calibrated model allows us to track the effects through space, which makes it a useful tool for regional policy evaluation.

5.A Data

The dataset used in this paper concern the 51 US states in the year 1997. The complete set can be found on the internet, at `http://knaap.com/gdata`. Data was collected from a variety of sources. We list them here, together with a download address.

- **Gross State Product.** Supplied by the Bureau of Economic Analysis, US Department of Commerce. The June 7, 1999 edition of the current-dollar GSP estimates were used. Available at `http://www.bea.doc.gov/bea/regional/gsp/`.

- **Employment.** Total nonfarm employment per state, from the Bureau of Labor Statistics. Available at `http://146.142.4.24/cgi-bin/srgate`. Request series SASxx00000000001, where xx is the state number.

- **Wages.** Average annual pay for 1997, from the Bureau of Labor Statistics, December 15, 1999 edition. Available at `http://stats.bls.gov/news.release/annpay.t01.htm`.

- **Interstate flow of commodities.** Bureau of Transportation Statistics 1997 State-to-state commodity flows in millions of US$. Available at `http://www.bts.gov/cfs/cfs97od.html`.

- **Distance between states.** Duncan Black kindly supplied a computer file with the latitude and longitude of each US county. I averaged these into state coordinates, weighing them with county employment. The distance between two states is then computed in miles using the great circle formula. For the distance within a state, I obtained the state area A_i and computed the quasi-radius as $\sqrt{A_i/\pi}$. This number approximates the average distance travelled within a state. State areas may be found at `http://www.census.gov/population/censusdata/90den_stco.txt`.

- **Factor shares.** The numbers in table 5.7 come from the BLS website at `http://www.bls.gov` and have ID numbers MPU300013 through MPU300017.

- **Weather data.** National Climatic Data Center, Asheville, NC. Tables can be accessed via `http://ols.nndc.noaa.gov/plolstore/plsql/olstore.prodspecific?prodnum=C00095-PUB-A0001`.

- **Mining.** Data from Smith (1997) available at `http://minerals.er.usgs.gov/minerals/pubs/commodity/statistical_summary/871497.pdf`

$\log(X_{rs})$	World 1	US 1	World 2	US 2	World 3	US 3	US 4	US 5
Obs.	10100	2601	8079	2201	10100	2601	2091	2042
Year	1994	1997	1994	1997	1994	1997	1997	1997
Estimation	OLS	OLS	OLS	OLS	Tobit	Tobit	OLS	OLS
$\log(\mathrm{dist}_{rs})$	-1.538	-1.181	-1.353	-1.044	-1.738	-1.330	-0.983	-0.987
	$[-0.041]$	$[-0.056]$	$[-0.032]$	$[-0.025]$	$[-0.043]$	$[-0.063]$	$[-0.024]$	$[-0.023]$
bord_{rs}	0.976	0.774	1.042	0.492	0.917	0.658	0.554	0.554
	[0.195]	[0.126]	[0.141]	[0.052]	[0.179]	[0.140]	[0.049]	[0.048]
own_{rs}	-	2.462	-	2.210	-	2.335	2.232	-
		[0.232]		[0.095]		[0.257]	[0.090]	
R^2	0.789	0.779	0.786	0.921	-	-	0.924	0.921
log L	-	-	-	-	-20306	-4422	-	-

World columns are from Table 1 in Redding and Venables (2001), *US* columns are own computations. Estimation 1 uses the full sample, including zeros. Estimation 2 uses a censored sample, from which the zeros have been eliminated. Estimation 3 again uses the full sample, taking care of the left-censored observations by using a Tobit estimation. Estimation 4 uses only the contiguous states, eliminating Hawaii and Alaska, as well as the District of Columbia. Estimation 5, finally, uses that sample without the within-state flows.

Table 5.2: Panel estimates for the gravity trade equation

$\log(w_r)$	World	US	World	US	World	US
Obs.	101	48	101	48	101	47
Year	1996	1999	1996	1999	1996	1999
$\log(\mathrm{FMA}_r)$	0.476	0.133	-	-	0.316	0.066
	[0.076]	[0.082]			[0.088]	[0.044]
$\log(\mathrm{MA}_r)$	-	-	0.479	0.257	-	-
			[0.063]	[0.029]		
$\log(\mathrm{DMA}_r)$	-	-	-	-	0.141	0.119
					[0.059]	[0.014]
R^2	0.346	0.079	0.610	0.601	0.584	0.613
Moran's I		0.197		0.317		0.404
$1 - F(I)$		0.0138		0.0006		0.0000

World columns are from Table 2 in Redding and Venables (2001), *US* columns are own computations. The dependent variable in World columns is GDP per capita. Bootstrapped standard errors are in parentheses (200 replications). First stage estimation is Tobit for the World columns, US 4 (see table 5.2) for US columns. Moran's I is computed on the residuals of the estimation, using a matrix of border-dummies as a weighing matrix. On the line below is the position of the statistic in a bootstrapped distribution function (100,000 replications).

Table 5.3: Market Access and wage levels

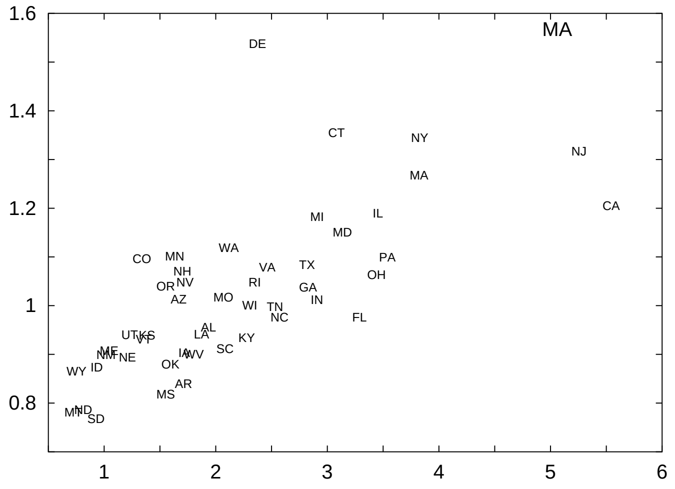

Figure 5.2: Predicted Market Access (horizontal, based on 1997 data) versus log wages (vertical, data from 1999, wages in ten thousands of dollars) for 48 states. MA regressors come from the US 4 estimation.

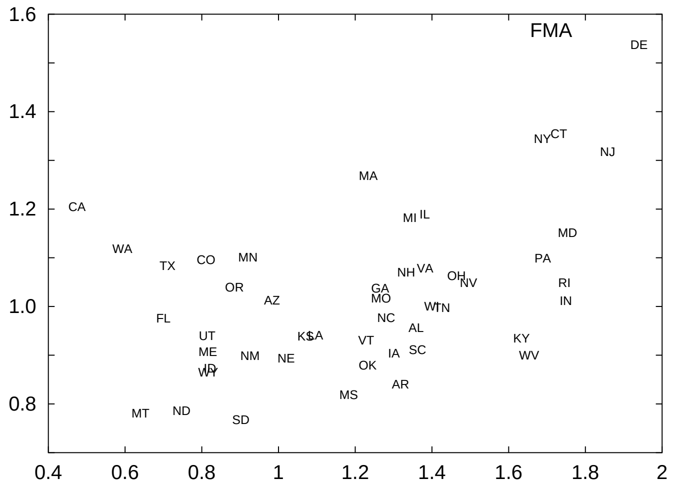

Figure 5.3: Predicted Foreign Market Access (horizontal) versus log wages (vertical) for 48 states. FMA regressors come from the US 4 estimation.

$\log(w_r)$	World	US	World	US
Obs.	101	48	101	48
Year	1996	1999	1996	1999
$\log(\mathrm{FSA}_r)$	0.532	0.118	-	-
	$[0.114]$	$[0.082]$		
$\log(\mathrm{SA}_r)$	-	-	0.345	0.229
			$[0.032]$	$[0.030]$
R^2	0.377	0.075	0.687	0.542
Moran's I		0.217		0.322
$1 - F(I)$		0.0091		0.0006

World columns are from Table 9 in Redding and Venables (2001), *US* columns are own computations. See the note under table 5.3.

Table 5.4: Supplier Access and wage levels

$\log(w_r)$	US	US	US	US
Obs.	48	48	48	47
Year	1999	1999	1999	1999
$\log(\mathrm{FMA}_r)$		0.042		0.044
		[0.063]		[0.046]
$\log(\mathrm{MA}_r)$			0.234	
			[0.041]	
$\log(\mathrm{DMA}_r)$				0.112
				[0.019]
nrmcdd	−0.103	−0.103	−0.049	−0.063
	[0.044]	[0.049]	[0.036]	[0.032]
nrmhdd	−0.010	−0.010	0.013	0.009
	[0.016]	[0.016]	[0.013]	[0.011]
minerals	0.022	0.021	0.006	0.005
	[0.004]	[0.004]	[0.005]	[0.005]
port	0.130	0.130	0.070	0.044
	[0.035]	[0.036]	[0.029]	[0.031]
R^2	0.545	0.551	0.756	0.776
Moran's I	0.164	0.128	0.230	0.205
$1 - F(I)$	0.0281	0.0586	0.0061	0.0127

Standard errors in parentheses. Except for the first column, these errors come from bootstrap methods (200 replications). First stage estimation for market access variables is US 4 (see table 5.2). Moran's I is computed on residuals, using a matrix of border-dummies. The position of the statistic in a bootstrapped distribution function is indicated below (100,000 replications).

Table 5.5: Exogenous amenities, Market Access and wage levels

$\log(w_r)$	US	US	US	US
Obs.	48	48	48	48
Year	1999	1999	1999	1999
$\log(\text{FMA}_r)$	0.169		0.076	
	[0.101]		[0.079]	
$\log(\text{MA}_r)$		0.232		0.144
		[0.107]		[0.111]
nrmcdd			-0.103	-0.070
			[0.057]	[0.046]
nrmhdd			-0.011	0.004
			[0.019]	[0.017]
minerals			0.020	0.012
			[0.005]	[0.009]
port			0.129	0.093
			[0.036]	[0.037]
R^2	0.073	0.599	0.547	0.724
Moran's I	0.163	0.291	0.105	0.145
$1 - F(I)$	0.0299	0.0013	0.0908	0.0423

Instrumental variables estimation. In the first two columns, instruments are the distance from New York and the distance from Los Angeles. In the third and fourth column, the four exogenous regressors are added to the set. Standard errors come from bootstrap methods (200 replications). First stage estimation for market access variables is US 4 (see table 5.2). Moran's I is computed on residuals, using a matrix of border-dummies. The position of the statistic in a bootstrapped distribution function is indicated below (100,000 replications).

Table 5.6: Instrumental variables estimation

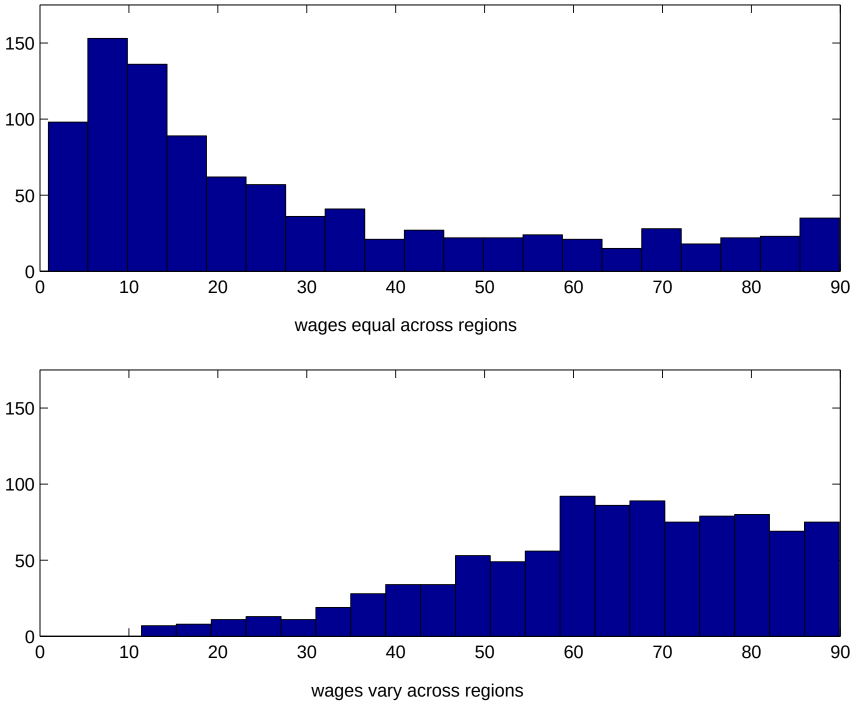

Figure 5.4: Distribution of the angle between the two vectors $\mathrm{d}T/\mathrm{d}\sigma$ and $\mathrm{d}T/\mathrm{d}\tau$ where T is the 9 by 1 vector of equilibrium trade flows between three regions. The top panel has wages fixed at one, the bottom panel has a random variation in wages between regions. An angle of $90°$ implies orthogonality of the two regressors, and an angle of $0°$ implies perfect colinearity.

	1997 factor shares costs ($ billions)	share (total)	share(relative)
Labor	934.3	37.8%	60.5%
Capital	516.7	20.9%	
Energy	63.8	2.6%	
Materials	610.5	24.7%	39.5%
Business Services	346.4	14.0%	

Data come from the Bureau of Labor Statistics and give the cost of different factors for the manufacturing sector in 1997. See also the appendix on data on page 144.

Table 5.7: Factor shares in US production, 1997

$P(\sigma, \tau, \delta)$	Import		Export	
Obs.	2043		2043	
Year	1997		1997	
σ	4.110	$[1.066]$	5.120	$[2.499]$
τ	$5.124 \cdot 10^{-4}$	$[2.199 \cdot 10^{-4}]$	$1.332 \cdot 10^{-4}$	$[8.957 \cdot 10^{-5}]$
δ	-0.235	$[0.106]$	-0.300	$[0.179]$
Parameters comparable to those in table 5.2.				
$\log(\text{dist}_{rs})$	-1.616		-0.547	
bord_{rs}	0.731		1.234	
R^2	0.725		0.780	

Estimation results from general equilibrium estimation as outlined in section 5.5.2. σ is the elasticity of substitution, τ and δ come from the transport cost function $T_{i,j} = \exp(\tau \text{dist}_{i,j} + \delta \text{border}_{i,j})$. Standard errors are Monte Carlo estimates based on 281 replications.

Table 5.8: General equilibrium estimation results

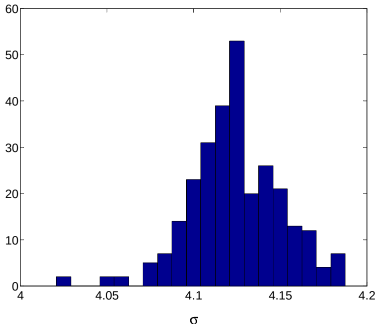

Figure 5.5: MC distribution of $\hat{\sigma}$, 'export' estimation, $1/40$ errors

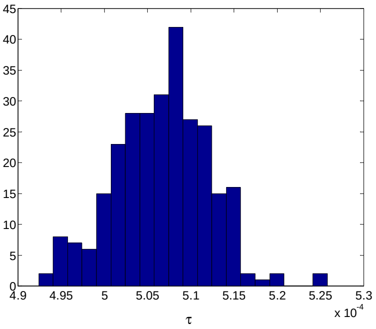

Figure 5.6: MC distribution of $\hat{\tau}$, 'export' estimation, $1/40$ errors

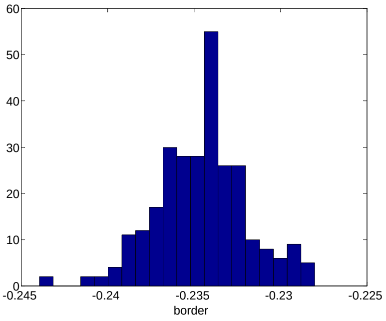

Figure 5.7: MC distribution of $\hat{\delta}$, 'export' estimation, $1/40$ errors

State	Demand change	Change w/o E	Change in n	Change in G	Change in p	Demand from FL
IL	14.6%	15.4%	−3.8%	−0.8%	−6.4%	16.5%
Neighbors						
IN	−1.8%	−0.4%	0.4%	−1.1%	−0.4%	0.3%
IA	−1.7%	−0.4%	0.5%	−1.2%	−0.5%	0.4%
WI	−1.7%	−0.3%	0.5%	−1.3%	−0.5%	0.6%
MO	−1.6%	−0.5%	0.4%	−1.1%	−0.4%	0.2%
Intermediate states						
PA	−1.5%	−1.1%	0.2%	−0.5%	−0.2%	−0.7%
CO	−1.1%	−0.9%	0.2%	−0.6%	−0.2%	−0.5%
TX	−1.3%	−1.0%	0.2%	−0.6%	−0.2%	−0.6%
Faraway states						
CA	−0.9%	−0.8%	0.2%	−0.4%	−0.2%	−0.8%
ME	−1.5%	−1.1%	0.2%	−0.5%	−0.2%	−0.6%

The first column indicates the state. The change in demand that each state experiences is in the second column; the third column shows what the change would have been without the accompanying drop in Illinois expenditures. Changes in the number of firms n, price index G and price p are next. The final column gives the change in demand from Florida for each state.

Table 5.9: Effects of a 10% decrease in Illinois wages

State	Demand change	Change in n	Change in G	Change in p	Demand from FL
IL	2.3%	0.3%	−0.8%	−0.3%	1.09%
IN	2.4%	0.3%	−0.7%	−0.3%	1.06%
Neighbors					
IA	−0.4%	0.0%	−0.1%	0.0%	0.01%
WI	−0.4%	0.0%	−0.1%	0.0%	0.02%
MO	−0.4%	0.0%	−0.1%	0.0%	−0.01%
KY	−0.4%	0.0%	−0.1%	0.0%	0.00%
MI	−0.4%	0.0%	−0.1%	0.0%	−0.01%
OH	−0.3%	0.0%	−0.1%	0.0%	−0.03%
Intermediate states					
GA	−0.2%	0.0%	−0.1%	0.0%	−0.08%
CO	−0.2%	0.0%	−0.1%	0.0%	−0.07%
TX	−0.2%	0.0%	−0.1%	0.0%	−0.07%
Faraway states					
CA	−0.1%	0.0%	0.0%	0.0%	−0.10%
ME	−0.2%	0.1%	−0.1%	0.0%	−0.07%

The first column indicates the state. The change in demand that each state experiences is in the second column; Changes in the number of firms n, price index G and price p are next. The final column gives the change in demand from Florida for each state.

Table 5.10: Effects of a 50% decrease in of transport costs between Illinois and Indiana

Chapter 6

An economic geography approach to evaluating a new Dutch railway link

6.1 Introduction

In May of the year 2000, the Dutch government commissioned a research project into the economic effects of a major infrastructural project involving the construction of a rail link between Amsterdam and the north of the country. The region around Amsterdam[1] is clearly the economic center of the Netherlands: 42% of employment and more than 48% of GDP is produced on an area that is no more than 15% of the country's total. Quite the opposite, the north of the Netherlands[2] is considered an economic laggard with 11% of GDP and 8% of employment on 27% of the country's area. Attempts to jump-start the northern economy with large subsidies and the forced relocation of government-owned companies so far have failed. The unfulfilled potential is aptly illustrated by relatively high unemployment rates and the swift return of the once-relocated companies to the West.

The construction of a government-sponsored rail link between the West and the North is thought to remedy this problem. Arguments in favor of construction center around the indirect effects of such a link. Of these there are two, external to the train's operator, that could benefit the country as a whole. The first effect concerns the commuting behavior of workers in the

---

[1] The statistics pertain to what is called the 'Randstad,' comprising metro Amsterdam, the province of Utrecht and the province of South Holland including metro Rotterdam. We will use the terms *Randstad*, *the West* and *the Economic Center* interchangeably in this paper. GDP statistics come from RuG/CBS (1999) p. 16, others from the 1998 LISA database.

[2] Although the north of the Netherlands is quite diverse, the term is usually reserved for the combination of the provinces Friesland, Groningen and Drenthe. These provinces have allied themselves in a bid for economic support from the central government.

West. With a fast rail link, they could relocate to the North while keeping their current jobs. The incentive to move comes from the relatively low prices of northern real estate as construction in the West reaches its natural limit. The decrease in pressure on the Western housing market would be considered a national benefit. This effect is explored in a parallel paper contained in Oosterhaven et al. (2000).

The second effect concerns the changes in economic activity that can be brought about by the link. Many companies that start in the North eventually move to the West, quoting their desire to be close to other companies (mostly those that deliver services, like advertising agencies and legal firms) as a reason for leaving. This desire is strong enough to overcome the higher prices of property, the tighter labor market and greater congestion of the West. With a rail link in place, the price of these services to firms located in the North would be lower, possibly shifting the balance in favor of location outside the center. Such induced activity is seen as the key to further economic development.

In this paper, we will explore the second effect using a Computable General Equilibrium (CGE) model of the Netherlands. The model fits in the New Economic Geography line of research and builds on a concept introduced by Venables (1996b). The basic structure resembles a similar model developed for the European Union by Bröcker (1999). The model is more detailed, however, introducing fourteen different sectors and their input-output linkages.

The effects that are captured by this kind of model lend themselves to an explanation in terms of forward and backward linkages (Hirschman 1958). A reduction in transport costs generally leads to a lower price of products consumed far from their producer. This leads to an increase in demand for this producer, which is an example of a backward linkage. Because the cheaper product can again be used as an intermediate input by local producers, they in turn can reduce *their* prices: this is an example of a forward linkage. Because we use local IO tables and detailed information about the effect of the rail link on travel times, we can track effects through the economy and derive detailed welfare effects.

An important part of the data that is used in the construction of this model is taken from the bi-regional input-output tables (RUG/CBS 1999) in which the economic ties between Dutch provinces are detailed. Some of the model's parameters come directly from this publication; others are estimated by fitting trade flows predicted by the model to flows observed in the tables. Special care is taken with respect to the spatial structure of the model: we discern between the transport of goods and people and in the latter case discuss the role of public transport.

This chapter continues as follows: in Section 6.2, we discuss relevant economic theory and the specification of the model. Estimation of the parameters and calibration of the model is done in Section 6.3. Section 6.4

discusses the project alternatives and shows the results of simulations. We evaluate the model in Section 6.5, where we attempt to quantify the uncertainty of the analysis and point out obvious weaknesses. Section 6.6 concludes.

6.2 The Model

There exist several models that explain spatial patterns of production by increasing returns to scale and positive transport costs. In these models, agglomeration is caused by the desire to overcome transport costs when selling ones product or making purchases. This similar desire on the side of producers and consumers leads to a feedback loop, resulting in self-enforcing agglomeration. The precise form of this loop differs between models.

We use a specification introduced by Venables (1996b) where firms use both labor and intermediate goods in production. Workers are not allowed to relocate, but firms enter and leave the market according to profitability. For some parameter values, a situation where most activity is located in one place is stable: the attraction for firms consists of the low price of intermediate goods and is self-enforcing.

We modify this model on several counts. Different sectors are introduced, leading to a richer set of possible outcomes (the effects of different sectors in these models are explored in Knaap, 2000). The labor market is simplified, so that it is in line with Dutch reality. Transport costs are differentiated according to what exactly is transported.

In the following sections, we use the convention that there are fourteen sectors indexed by s. The country can be divided in fourteen major regions (twelve provinces and two metropolitan areas) indexed by p, in forty COROP regions indexed by o or in 548 communities indexed by c. A full description of this convention, and the available data on each level, can be found in Appendix 6.A.

6.2.1 Production and Utility

Specification

Utility of a representative consumer in province p is given by

$$U_{i,p} = \prod_{s=1}^{14} U_{i,sp}^{\theta_{sp}} \tag{6.1}$$

where

$$U_{i,sp} = \left(\sum_{c=1}^{548} n_{cs} X_{i,cs}^{1-1/\sigma_s} \right)^{\frac{1}{1-1/\sigma_s}} . \tag{6.2}$$

with X_{cs} the level of consumption by person i of a sector s-product from community c. As it turns out that all firms from a certain community in the same sector use the same price, the number X_{cs} holds for all those firms. The number of firms in community c that are in sector s is given by n_{cs}.

As seen above, utility is computed in two stages: first, according to (6.2), sub-utility within each sector is computed by aggregating purchases from all communities. This aggregation is done by a CES function, indicating that the firms within a sector are in a state of monopolistic competition (Dixit and Stiglitz 1977). The size of the different $X_{i,cs}$'s depends on the price of the product and the sector-specific elasticity of substitution σ_s. Sectorial utilities $U_{i,sp}$ are combined using the Cobb-Douglas function (6.1). This specification implies that each sector receives a fixed share of the consumers' budget.

As appears from this specification, we allow for different utility functions in different parts of the country: each major region has its own set of utility parameters Θ_p. While it is unclear if regional peculiarities of this kind are a stable phenomenon, this specification allows us to take parameters Θ_p directly from the bi-regional IO tables: they are simply the share of the consumer budget devoted each sector. Assuming that each region described in the tables has its own preferences is a convenient short cut that nonetheless has its price: a higher share of the budget devoted to a certain sector could also indicate a lower price of those products in a certain region, possibly due to transport costs. While recognizing this problem, we employ the specification in (6.1) for the sake of simplicity.[3]

On the production side, we assume that firms face the following production function:

$$Y_{j,ps} = c_{ps} L_j^{\alpha_{ps}} Q_j^{1-\alpha_{ps}} \tag{6.3}$$

$$Q_{j,ps} = c'_{ps} \prod_{s'=1}^{14} Q_{j,s'}^{\gamma_{p,s,s'}} \tag{6.4}$$

The production of any firm j is thus a Cobb-Douglas aggregate of used labor L_j and intermediate goods Q_j. The parameter α_{ps} of function (6.3) varies per sector and per major region. It is computed from the bi-regional IO tables: $1 - a_{ps}$ is the share of intermediate products used in production. The sector-region specific constants c_{ps} and c'_{ps} allow us to use a simple form for the cost function later on. Derivations of these constants are in appendix 6.B.1.

The intermediate good itself is an aggregate of goods and services from all fourteen sectors, as is shown in formula (6.4). Once again the aggregation is of the Cobb-Douglas variety, with parameters $\gamma_{p,s,s'}$ taken directly from the IO tables.

---

[3] The variance of θ_{sp} over p around the average θ_s was typically around 5% of the θ_s.

On the sectorial level, we assume monopolistic competition. So, while the input share of a certain sector may be a constant γ, the actual producer that is chosen to supply the input is dependent on the price. This is a very appealing assumption, as in reality parameters like γ are often dictated by technical constraints, but within these constraints the producer is free to shop around for the cheapest supplier. The specification of $Q_{j,s}$ is identical to that of sectorial sub-utility in formula (6.2):

$$Q_{j,s} = \left(\sum_{c=1}^{548} n_{cs} X_{j,cs}^{1-1/\sigma_s} \right)^{\frac{1}{1-1/\sigma_s}} . \tag{6.5}$$

In this specification, it is essential that producers and consumers share the same elasticity of substitution σ_s. This way, the demand curve from both parties is identical and the optimal price for the supplier is the same, regardless of the type of customer. It also implies that we can use the same price index for both producers and consumers. Different values of σ_s would make the model much more complicated and are not considered in this paper.

Solution

The standard Monopolistic Competition results (MC hereafter) hold in this model, leading to familiar, if somewhat elaborate, expressions for demand and supply. Consumers and producers both exercise demand. If we look at a consumer in community c, major region p with income w, her demand for a certain product from producer j in sector s', located in community c' will be

$$D\left(p_{j,c',c,s'}\right) = w \cdot \theta_{s',p} \cdot \frac{p_{j,c',c,s'}^{-\sigma_{s'}}}{G_{c,s'}^{1-\sigma_{s'}}} \tag{6.6}$$

with the price index defined by

$$G_{c,s'} = \left(\sum_{c''=1}^{548} n_{c'',s'} \cdot p_{j,c'',c,s'}^{1-\sigma_{s'}} \right)^{\frac{1}{1-\sigma_{s'}}} . \tag{6.7}$$

Similarly, demand from a producer in sector s, community c which is in major region p, who spends wL on labor, will demand from producer j

$$D\left(p_{j,c',c,s'}\right) = wL \cdot \frac{1-\alpha_{p,s}}{\alpha_{p,s}} \cdot \gamma_{p,s,s'} \cdot \frac{p_{j,c',c,s'}^{-\sigma_{s'}}}{G_{c,s'}^{1-\sigma_{s'}}} .$$

Notice that, as usually in MC models, a positive quantity is demanded from *each* producer, no matter how high the price (and no matter how far away). This may cause a problem later on as many products in real life are not suitable for transport over long distances.[4] We will return to this problem in Section 6.3.2.

6.2.2 The Labor Market

One of things that sets this model apart from that in Venables (1996b) is the specification of the labor market. Usually a completely inelastic labor supply is assumed, where a given amount of labor is always employed and the wage is computed as the closing variable of the model. Wage differences lead to marginal cost changes and to price differences between regions, which is an important step in the model's final results.

We feel that such a competitive wage-setting environment does not accurately reflect the situation in the Netherlands. Wages are negotiated on a national level and thus are not different between regions. This has several repercussions: first, there are no incentives to migrate between regions in order to receive a higher nominal wage. Second, local labor markets do not always clear. Unemployment is prevalent in those regions where excess labor supply exists. The unemployed nonetheless are able to exert demand, the same way the employed workers do, through unemployment benefits.

Therefore, the labor market is modelled as quantity-oriented and demand controlled. We assume that the wage is equal throughout the country and set it to 1. Any shocks in labor demand are absorbed by hiring or firing workers, implicitly assuming that there are no constraints in labor supply: each community has a sufficiently large pool of unemployed to use in times of increased labor demand.

We model the effect of income taxes and unemployment benefits in an even cruder way: all incomes are taxed at a rate of 100% and then redistributed to all inhabitants. This implies that the consumer income in any community is proportional to the number of inhabitants. Because we do not model the labor supply decision, this rather unorthodox taxation scheme does not have an impact on the supply of labor.

6.2.3 Transport Costs and Prices

Specification

It is customary in models of this kind to let transport costs take the form of leakage: a certain fraction of the transported product is lost along the way, the size of the fraction determined by the distance travelled. By incurring

---

[4] The result that a positive quantity is demanded from each producer regardless of the price (indicating an indispensability of each product variety) does not necessarily follow from the assumptions underlying monopolistic competition, but rather from the iso-elastic demand function that we have assumed. It occurs usually in MC models the sense that this type of demand function is usually employed.

transport costs in the product itself, there is no need to explicitly model a transport sector and prices can easily be adjusted for distance.

We modify this *iceberg approach* to account for the fact that there are two types of transport, and the new infrastructure will change only one of those types. Transport of goods is assumed to be unaffected by the new link, as it takes place mostly by truck, ship or pipeline. Passenger transport on the other hand, the second type, is definitely affected by the new link.

We compute transport costs as follows: in general, transport causes a markup on the price of a product, depending on the distance d that is travelled, equal to

$$f(d) = 1 + \nu d^{\omega}. \tag{6.8}$$

Depending on the sector s to which the product belongs, a share π_s is goods transport and a share $1 - \pi_s$ uses passenger transport. The total transport markup thus is equal to

$$\begin{aligned} \tau_s(d) &= [f_g(d)]^{\pi_s} \cdot [f_p(d)]^{1-\pi_s} \\ &= [1 + \nu_g \cdot d^{\omega_g}]^{\pi_s} \cdot [1 + \nu_p \cdot d^{\omega_p}]^{1-\pi_s} \end{aligned} \tag{6.9}$$

Distance for goods transport has been measured in kilometers. The distance for passenger transport is measured in minutes and is computed an average between public transport time and driving time (for these two variables, see also Section 6.A.2). The parameters ν_i and ω_i are estimated in Section 6.3. Parameters π_s have been obtained exogenously and are specified in Section 6.A.3.

Solution

The marginal costs for firm j, which is in sector s, community c and region p are equal to

$$\mathrm{MC}_{s,c,p} = w^{\alpha_{ps}} \cdot \tilde{G}_{cs}^{1-\alpha_{ps}} \tag{6.10}$$

where the price index of intermediate goods $\tilde{G}_{cs}$ is defined as

$$\tilde{G}_{cs} = \prod_{s'=1}^{14} G_{c,s'}^{\gamma_{p,s,s'}}$$

with the price index for sector s in community c, $G_{c,s}$, defined above in formula (6.7).

The optimal price for the above firm is, as usual in MC, a markup times the marginal costs:

$$p_{j,s,c} = \frac{\sigma_s}{(\sigma_s - 1)} \mathrm{MC}_{s,c,p} \tag{6.11}$$

This gives the price $p_{j,s,c}$ which holds in community c in which the firm operates. The price in another community c' is found using the specification of transport costs above:

$$p_{j,c',c,s} = \tau_s \left(d_{c,c'}\right) \cdot p_{j,s,c} \tag{6.12}$$

where $d_{c,c'}$ is the distance between the two communities.

6.2.4 Computation

The actual computation of the price that a company charges for its product in a given community is fairly complicated. As follows from (6.10), each price is a function of local wage (which is zero, by definition) and the local price index of intermediate goods. This price index depends on the price of nearly every other available good in the country, as well as transport costs for all these goods. In turn, these prices each depend on all other prices and applicable transport costs. The equations that describe these pricing decisions[5] cannot be solved analytically. In practice, a numerical procedure is used where all prices are set to one and the system is allowed to iterate until convergence. This presupposes knowledge of the parameters ν_i, ω_i and σ_s. Those parameters have been estimated - that procedure is detailed in Section 6.3. It also presupposes knowledge of the number of firms $n_{c,s}$ in each sector in each community. As it turns out, the number of firms in a community is proportional to the product of α_{ps} and $L^{total}_{c,s}$. The latter is the amount of labor that is used by the sector in that community, both are a known variables. The proof of this is deferred to appendix 6.B.2.

6.3 Estimation and Calibration

6.3.1 Procedure and Data

In the previous section, we specified the model and pointed out where the data came from and how we found the model's parameters directly from the IO tables. That is, up to eighteen unknown parameters that we will estimate in this section. We will call this set of parameters Γ, and they are

- the fourteen elasticities of substitution σ_s, each particular to a specific sector.
- the parameters of the two transport cost functions, ν_i and ω_i ($i \in \{g, p\}$).

We will estimate these parameters in the following way: for any given set of values for the unknown parameters, we can use the model to compute the demand that is exercised by each region upon every producer in

---

[5] In effect, formulas (6.7) and (6.10) through (6.12).

the country. Adding a subset of these numbers in an appropriate way, we can compute any flow of trade in the country, given that we have a set of parameters Γ.

For instance, we can compute the demand for agricultural goods exercised by consumers and businesses *inside* the province of Utrecht, upon suppliers that are located *outside* this province. This would be an intensive computation: we would have to establish prices for all goods, compute the amount of money allocated to agricultural goods in each Utrecht community, and use prices and transport costs to divide this budget between agricultural producers in all 548 communities. Then we would have to compute how many goods each budget will buy, add up all the goods bought outside the province, and sum over all buying agents inside Utrecht. This is of course a tedious task, which can fortunately be left to the software implementation of the model. Once it is done, we have replicated two entries from the bi-regional IO table of Utrecht: the flow of agricultural goods from outside the province into the province, used as intermediate good and used as consumption good.

The number of trade flows that can be lifted from the IO tables is, in principle, equal to $11,760$. There are fourteen bi-regional tables; each gives four sets of flows: *from* inside and outside the region *to* inside and outside the region. Each set consists of a fourteen by fourteen matrix with flows of intermediate goods and fourteen flows of final goods.[6]

We will estimate the parameters by trying different sets of Γ and finding the one that minimizes the sum of squared differences between the predicted (log-)flows of goods and the (log-)flows in the IO tables. The logarithms are used so that the larger flows do not dominate the estimation. For several reasons, we do not use all the available data, both discarding some flows and summing to aggregates.

First of all, the data regarding flows *from* outside a region *to* outside a region are discarded. The reason for this is that the quality of the data is thought to be poor; it is constructed by as a rest-category subtracting the other flows from a total. This also creates a redundancy in the data, as the same totals are used again and again.

We aggregate the rest of a data because of two reasons: first of all, computational restrictions limit the number of data points that can be digested in a reasonable time. Because we need to find an optimum over eighteen parameters, the computation involving a single set G cannot take too long. Secondly, many flows are insignificant and measured with a large error. For example, the mining sector in the reclaimed-land province of Flevoland is virtually non-existent, leading to very small trade flows which contain no useful information. Using this data in logarithms would contaminate our set.

---

[6] To make matters confusing, there are fourteen regions and fourteen sectors.

So, we sum observations until we are left with 588 datapoints: for each of 14 regions we have the sales of each of 14 sectors, to customers inside the own region (both final and intermediate) and to customers outside the region. For each of 14 regions, we also have 14 values giving the purchases from outside the region per sector. We will match these 588 datapoints with those predicted by the model.

6.3.2 Problems with the estimation

Transportability

Within the model as we have described it so far, the consumption decision is based on prices, which in turn are the result of transport costs. A strong preference for local goods thus can only be explained by very high transport costs. As it turns out, for some sectors the preference for local goods is so strong that it would imply incredibly high costs of transport. For instance, in our measured year consumers in the province of Utrecht spent 1749 mln guilders on education. Almost 92% of these expenditures were made *within* the own region, which produces only 9% of the country's educational output. Clearly, if price was the only issue, transport costs would be immense. There seems to be something in the nature of education that makes in less suited for trade.

So far we assumed in our model that there are no non-tradeables. To allow for these, we introduce sector-specific parameters that measure *transportability*. Obtained from the regional balances of goods (RUG/CBS 1999), these parameters indicate the degree to which output can only be produced on the spot, because it is over-the-counter or personal. The parameters are exogenous to the model and given in appendix 6.A.3. In our model, we account for transportability by dividing expenditures on each sector into expenditures on non-tradeables and expenditures on tradeables using the new parameters.

Identification

In related work, Bröcker (1999) found that the estimation of the parameter set Γ suffers from the problem that the parameters may not be separately identifiable. He proves that for a slightly different transport cost function, which nonetheless resembles the one in (6.8), parameters are not identified at all.

The problem is quite intuitive: increasing the costs of transport or increasing the elasticity of substitution has exactly the same effect: more local goods will be consumed, either because imported goods become more expensive or because the (cheaper) local goods can be more easily substituted for (more expensive) imported goods. This leads Bröcker to use extra data

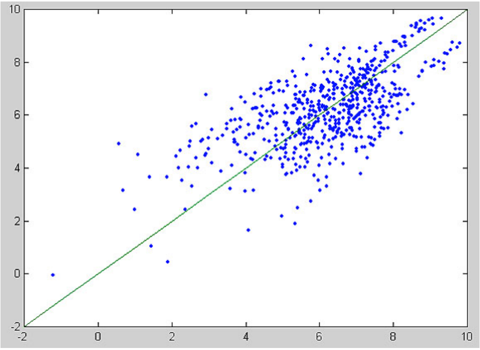

Figure 6.1: This plot shows the ability of our model to predict the flows of trade. Each point represents a flow—the x-coordinate is the predicted flow, and the y-coordinate is the actual flow. Both are in logs. Perfect prediction would mean that all points lie on the $45°$ line, which is also drawn.

about the importance of transport costs in final good prices in his estimations.

We have found that our model suffers from the same problem, in the sense that the search for optimal parameters seems to take place in a lower dimensional subspace of the parameter space. Nonetheless, we found an optimal constellation of parameters which were within reasonable bounds, given other estimations of this kind (for instance Bröcker 1999, Hanson 1998), for instance).

6.3.3 Estimation results

The parameters that minimize the sum of squared errors are in table 6.1. The graph in figure 6.1 shows the goodness of fit of the estimation. The (log) flows of trade as predicted by the model are on the horizontal axis, the actual (log) flows of trade are on the vertical axis. a $45°$ line is drawn.

Parameter	Estimate	Parameter	Estimate	Parameter	Estimate
σ_1	11.1	σ_8	23.6	ν_g	0.006
σ_2	8.3	σ_9	8.2	ω_g	0.770
σ_3	12.5	σ_{10}	13.7	ν_p	0.010
σ_4	16.4	σ_{11}	14.2	ω_p	0.593
σ_5	24.0	σ_{12}	12.2		
σ_6	15.1	σ_{13}	12.5		
σ_7	29.8	σ_{14}	18.7		

Table 6.1: Estimates of the unknown parameters Γ as obtained from a non-linear least squares procedure

Many of the estimated parameters are much higher than those measured in chapter 5. This is not unusual; Hanson (1998), for instance, measures $\sigma = 10.4$ for United States counties. At face value, these results do seem to imply that many markets are close to perfectly competitive.

So far, we have used the wage as a numéraire. This means that all values inside the model, including the predicted flows of trade, are denoted in terms of w. During the estimation, when we must compare the predicted flows of trade with actual numbers denoted in millions of guilders, we must convert our internal figures. This is done by multiplying them by a factor so that, on average, the prediction is correct. This procedure is equivalent to estimating an intercept in figure 6.1. This factor, or the intercept, is in turn an indication for the actual value of our numéraire, the wage level w. We find an intercept of -2.55 which indicates an average wage of 77.9 thousand guilders. This is surprisingly close to the actual figure, which is 79.2 thousand for our measured year.

With the numbers in table 6.1, we can do some back-of-an-envelope calculations about the effect of distance on demand. We use a typical elasticity of substitution $\sigma = 12$ to compute at which distance half of wholesale is lost because of transport costs. For this distance $d_{1/2}$, there must hold that

$$\frac{1}{2} = \left(1 + \nu_i \cdot d_{1/2}^{\omega_i}\right)^{-(\sigma-1)}, i \in \{g, p\}. \tag{6.13}$$

Using this formula, we compute that a customer who would buy 1 guilder's worth when the supplier would live next door, buys exactly 50 cents' worth from the same supplier when the distance between them is 21 kilometers and all hauling involved is goods transport. Similarly, if all transport is of the passenger kind (*i.e.* the supplier must meet the customer in person) the potential wholesale is halved every 22 minutes.

6.3.4 Calibration

We now have the complete set of parameters, estimated using data from the 1990s. The aim of our model is to evaluate the impact of certain infrastructural changes in 2020. To be able to do this, we calibrate the model to a dataset that describes the situation in 2020 if none of the proposed projects is carried out. This dataset consists of the travel-times between all Dutch communities based on the projected state of public transport in 2020 (NEI/HCG 2000; the so-called *reference matrix*) and the projected number of jobs in each community in each sector in 2020 (TNO Inro 2000). We call this set the null-alternative; it is constructed exogenous to the model by based on economic scenarios from the CPB and known plans to upgrade infrastructure. More information about the different scenarios is in appendix 6.A.2.

Given the parameters, the amount of labor used in each sector in each community, and the matrix of transport times we can compute supply and demand for each sector in each community in 2020, in the null scenario. Unfortunately, given that the data in the null-scenario come from outside our model, there is nothing that guarantees the equality of supply and demand at this low level. Yet, to start out with a balanced model we must somehow equate the two. This prompts the introduction of a class of 'offset' variables, which take the value of demand minus supply for each sector in each community. We now redefine supply as the number computed with our model *plus* the appropriate offset variable.

One of the interesting figures concerning the implementation of a certain project will be the change in the number of jobs that ensues. Using the above patch, we can compute this number: the change in transport costs will give the change in demand; matching (the redefined) supply to this change in demand using formula (6.3) gives the needed amount of extra labor. The use of extra labor in turn triggers new demand effects that make their way through the economy.

6.4 Simulation Results

6.4.1 Some remarks

We evaluate five scenarios that are alternatives to the null-scenario to which we have calibrated our model above. The only difference between these scenarios and the null-scenario is in the matrix with travel-times. In each scenario, a different infrastructural project has been implemented and the changes in travel-time between each set of two communities has been computed. These computations are done outside our model by NEI/HCG (2000). Details are in section 6.A.2.

When we run our simulations, there are two more factors which we must take into account. First of all, while this study is concerned with the Netherlands as a closed country, we feel uncomfortable letting all the extra demand generated by the changes in the economy be absorbed by domestic producers. That is why we only let 50% of the extra demand come to bear on the Dutch market, letting the other half leak out of the country. As it turns out, this measure is does not have a large impact because the *extra* demand is very small; it is the distribution of demand that changes the most.

Secondly, we take account of a third sector-specific exogenous characteristic, *exogenous ties*. With this parameter, we incorporate the fact that for some companies, the choice of location and level of activity is wholly independent of prices, as they are tied to their location and their customers are tied to them. This can be because of localized natural resources or because of a fixed local clientèle, as in the case of municipal governments. The parameter gives the share of companies for which changes in price (as a result of changes in transport costs) does not alter their scale of operation, except for sector-wide changes in demand. For instance, the a municipal government which is 100% tied to its location and customers, does not sell any more or less because of changes in transport costs. However, if the changes lead to a smaller demand for *all* government services this does affect the municipal government. More on this parameter can be found in appendix 6.A.3.

6.4.2 Endogenous number of varieties

We encounter a severe problem during the first attempts to simulate a new equilibrium with the model. This problem is the result of our specification of the labor market, which was detailed in section 6.2.2. There, we assumed that there exists an infinite supply of labor at each location at the present wage. The labor market is demand-constrained and is characterized by positive unemployment in all locations. This assumption followed from the fact that the wage is not thought to be a regionally differentiated variable in the Netherlands.

However, as it turns out, the model has become unstable because of this assumption. The instability works as follows: if at a certain location demand is increased, the number of jobs and hence the number of firms goes up (formula 6.17). With the new firms, new varieties are introduced which lead to a new increase in demand and a new increase in the number of jobs. Because of the infinite supply of labor at unit cost, this process becomes explosive. Table 6.2 shows the dynamics of such a run. We evaluate the MZB scenarios and for two cities, we give the change in the number of jobs after each iteration of the model. While Almere, situated along the new line, grows explosively the more peripheral city of Eindhoven looses all its employment.

	1	2	3	4	5	6	7
Almere	2063	3757	5417	7305	9691	12931	17448
Eindhoven	−5	−43	−105	−188	−293	−423	−591

Table 6.2: The change in employment in two cities after a project, in seven iterations of the model. The number of firms varies endogenously.

To stop the model from exploding the way it does in table 6.2, we leave the number of firms fixed during simulation runs. This means that our analysis only picks up effects that relate to the redistribution of demand as a result of the new prices, and to the fact that less product is wasted in transport. Effects that come about because of an change in the number of varieties at certain locations are no longer part of the analysis.

6.4.3 Figures

We evaluate five possible scenarios involving infrastructural projects and compare them to the reference scenario in which only planned improvements to infrastructure are made. The scenarios are in appendix 6.A.2. For each scenario, we compute a new equilibrium of supply and demand, given the changed costs of transport. Each equilibrium is characterized in two ways:

1. We compute the change in the number of jobs in each of four regions. We take the two regions that are supposed to benefit most, the West and the North, the Flevo region through which most of the lines go, and the rest of the country. By aggregating the results for individual communities to this level, we hope to average out part of the error and show the result for regions that are politically interesting.

2. The infrastructural projects lead to a change in prices for each community, as services from far away get cheaper. Looking at the local price index allows us to compute a welfare effect of the project: the increase in utility that is achieved in a community by the lower local price index of consumption could also have been brought about by raising in municipal income with a certain amount of money. The sum of those equivalent increases in income give an indication of the welfare effect of the project.

	HIC	HHS	ZIC	MZB	MZM
North	634	1798	906	3496	3077
West	239	417	1185	2150	2503
Flevo	368	622	396	2075	2455
Rest	−1256	−2887	−2523	−7796	−8078

Table 6.3: Change in the number of jobs per region after each of five project alternatives

	HIC	HHS	ZIC	MZB	MZM
Δ CPI (in %)	−0.02	−0.06	−0.02	−0.09	−0.09
eqv. Δ GDP (mln95DFl)	141.6	344.7	124.1	577.5	553.5

Table 6.4: Change in the price index of consumption for the average consumer

The first results, the change in the number of jobs, are in table 6.3. We see that, as expected, the regions at the ends of the line are the biggest winners: both the North and the West gain the biggest number of jobs. Both regions experience an increase in demand (from each other) and a decrease in the price of intermediate goods, leading to higher order effects. If the schedule calls for frequent stops along the way, the middle province of Flevoland also shares in the gains. The rest of the country pays the price, though. Because the effect is largely redistributive (products from easier-to-reach regions become more popular vis-à-vis products from other places) the jobs that are gained in the North and the West are lost in the rest. Six maps with an indication of the geographical spread of these results are drawn in figure 6.2 on page 179.

The numbers in table 6.4 show that lower costs of transport lead to a lower price index of consumption for the average consumer. This is to be expected: since transport costs, which are a pure loss in the model, decrease everybody is better off. The assumption that we made about the redistribution of income implies that this is literally true for everyone, even those living in the regions that loose jobs. While the percentage-point figures in table 6.4 are not very impressive, the amount of money that is needed to create an increase in GDP with a comparable effect on utility is quite large.[7]

6.5 Evaluation

The study in this paper is part of a larger effort to gauge the most important effects of the infrastructural projects that are currently being proposed (Oosterhaven et al. 2000). The design of this effort is such that each sub-problem is analyzed in such a way that the effects from other sub-studies are deliberately left out. Then, all effects can be added up in the end with the risk of double-counting. It is therefore that we have not discussed such matters as migration by workers, international repercussions and the environmental impact of the projects. This sub-study has been limited to the economic redistribution that is to be expected after each of the projects.

In the course of this study, a large model has been constructed from the ground up, in a limited time. Some shortcuts had to be taken here and there, leading to some matters not getting the attention that they probably deserved. The exogenous sector-specific parameters in appendix 6.A.3 were picked by experts after consulting data on the subject, but not estimated rigorously. Due to the non-linear character of the model, it is hard to quantify the effects that errors in these parameters can have.

Secondly, the concept of an endogenous number of firms had to be abandoned after the model turned out to be unstable. This leads to an underestimation of the effects of a new link: any effects that we find with a

---

[7] Estimates are based on 1995 GNP of 640.56 billion DFl (CBS, Statline 2000).

fixed number of firms are sure to be larger when the variety-effect is taken into account.

However, this abandonment may be a blessing in disguise as it relieves us of another problem: during parameter estimation, the elasticity of substitution and the costs of transport were only weakly identified. With the results that we now present, these parameters operate together and need not have been estimated separately. A simulation involving the variety-effect would have used the elasticity of substitution intensively. Now, this elasticity is only used in conjunction with the costs of transport, except when the consumer price index is computed in section 6.4.3.

Thirdly, the use of *iceberg* transport costs, while theoretically convenient, leads to strange results here and there. For instance, a decrease in transport costs means that less product 'melts' along the way, which has a dampening effect on demand. This leads to an underestimation of the direct effects. The fact that one of our fourteen sectors is actually the transport sector makes matters even more awkward. A better way to include transport is the subject of further research.

6.6 Conclusion

We have constructed a spatial CGE model for 548 Dutch communities in 14 sectors, based on New Economic Geography principles. Our model can best be compared to the one in Bröcker (1999) and Venables and Gasiorek (1996) and uses intermediate products as in Venables (1996b). We calibrate the model to a base scenario for the year 2020 and use it to evaluate five infrastructural projects on which the Dutch government is about to decide.

We have chosen to model a demand-constrained labor market as if there is an infinite supply of labor available at a fixed wage. This leads to an instability in our model as the number of firms is derived from the amount of labor used. This problem is mitigated by assuming that the number of firms remains constant after a project has been implemented. Because of this, our results indicate mostly redistributive effects.

The most ambitious plan, called MZM in this paper (see appendix 6.A.2), leads to a shift of about 8,000 jobs. These jobs are gained in the North and the West, because of direct demand effects (each region's products are cheaper for the other) and indirect effects: because of cheaper intermediate products prices go down. The same plan leads to a decrease in the consumer price index. While this decrease is felt mostly in the North, all consumers benefit. The welfare increase that is the result of this decrease in prices is equivalent to one obtained after raising GNP about 550 mln DFl.

We have had to make a number of adaptations to the plain CGE model in order to get things to work properly. Sometimes, these adaptations are of an *ad hoc* nature and their effects on the are hard to gauge. Section 6.5

discusses the uncertainties that go along with this analysis.

Finally, the construction of this model can be seen as a first step toward the construction of a larger spatial CGE model of the Netherlands which can be used to help with infrastructural decisions in the future. TNO Inro and the University of Groningen are working on such a model at this time.

6.A Data and Conventions

6.A.1 Division of the Economy

In the Dutch economy, we identify fourteen sectors that produce goods and services. These goods and services are consumed by the public and demanded by other firms as intermediate products. The sectors are indexed by the variable s and are specified in Table 6.5.

Sector	Name	Sector	Name
1	Agriculture	8	Transport and Communications
2	Mining	9	Finance and Insurance
3	Industry	10	Other services incl. Real Estate
4	Public Utilities	11	Government
5	Construction	12	Education
6	Trade and Repairs	13	Health
7	Hotels, Restaurants, Bars	14	Culture and Recreation

Table 6.5: The fourteen different sectors in the economy

For each sector, we have used a number of coefficients concerning the use of different kinds of transport, tradeability and the nature of the exogenous ties to the location. These coefficients can be found in section 6.A.3.

Geographically, we divide the country into 14 major regions. The basis of this division are the twelve Dutch provinces. Ten of these are major regions, all except South Holland and North Holland which are each divided into two major regions: Metro Amsterdam including the area around the North Sea Canal is a major region, as is the rest of North Holland. Metro Rotterdam and the Ports is a major region, as is the rest of South Holland.

Each major region in turn consists of one or more COROP regions. There are 40 COROPs; the COROP division was originally intended to mark the size of local labor markets. Each COROP, finally, is divided into a number of communities. The model uses the 1998 situation regarding the municipal borders, leading to a total of 548 communities. The largest community is Amsterdam with 718,151 inhabitants, the smallest is the island of Schiermonnikoog with 1,003 inhabitants.

6.A.2 Available Data

IO Tables

Our most important source of data are the bi-regional input-output tables compiled in RuG/CBS (1999). Any reference to 'the IO tables' in this paper concerns this publication. The tables are available for twelve provinces and two metropolitan areas. Of each of these regions we know the internal 14x14 IO-table as well as external trade, summarized in two 14x14 IO tables (one for inputs from the rest of the country used and consumed locally and one for local outputs used and consumed in the rest of the country). From these tables, we derive not only the IO-structure on the regional level, but also flows of trade between the fourteen regions which help us estimate the model's parameters.

Community level data

We use the LISA (1998) database which gives, among other things, the amount of labor employed in each sector in each community in 1998. From this 548 by 14 matrix, we can derive production per sector per community and the number of firms per sector per community.

Distances

We discern two types of distance in this model. For goods transport, distance is measured in kilometers. We have computed the distance by car between all the possible pairs of communities using a CD ROM with travel information (AND 2000). This distance is used for goods transport in all scenarios, indicating that nothing changes with regard to this type of transport.

For passenger travel, we measure the distance between two communities in minutes. This distance is an average between the travel time by car (derived from AND (2000)) and by public transport. For the latter, we have a travel time matrix for detailing distances between any couple of communities for all scenarios (TNO 2000). This matrix is the one that changes most between scenarios, obviously. The weighing is done with *modal split* numbers supplied by the Netherlands Economic Institute (NEI 2000). For each scenario there is a modal split matrix for 28 areas, which takes into account the substitution effect that follows the construction of new infrastructure. This matrix is extrapolated to the 548x548 community pairs.

Scenarios

We use one economic scenario that is the basis for our calibration of the year 2020. This scenario gives the number of jobs in each sector in each community in 2020 and the number of inhabitants in each community in the same year. It is compiled by TNO Inro (TNO 2000) and is based on the CPB's 'European Cooperation' scenario and a separate regional model used at TNO Inro. In principle, we could use the 2020 scenario as a test on our model, calibrating it on the 1990s data en checking whether the 2020 scenario leads to an economy in equilibrium at the community level. When we do this, we find that the scenario leads to a severe disequilibrium. When investigated more closely, it turns out that the problems arise because of developments incorporated in the scenario that cannot possibly be predicted by our model like a shift towards service industries.

There are six travel time scenarios. We give them in table 6.6 below, referring the reader to the main report (Oosterhaven *et al*, 2000) for a more detailed description. The table gives a brief description of the project as well as the projected travel time between the northern city of Groningen and Schiphol Airport in the West.

Scenario	Description	Groningen - Schiphol
REF	The null alternative. This includes the Hanzelijn between Lelystad and Zwolle, which is yet to be constructed.	118
HIC	Hanzelijn + IC. The only difference with REF is that trains will go at a higher speed.	102
HHS	Hanzelijn - high speed. A high-speed train replaces the intercity service on the HIC scenario, calling at the larger stations.	71
ZIC	Zuiderzeelijn IC. New track is constructed between Lelystad and Drachten, leading to a straight link between the North and the West. The track is serviced with intercity trains.	89
MZB	Magnetic track. A new technology is used to create super-high speed trains which travel from Groningen to Amsterdam in a straight line. All trains call at all major stations.	59
MZM	Magnetic track - metro schedule. As MZB, but with a schedule that has non-stop trains between only a few terminals.	45

Table 6.6: Six scenarios for travel time in 2020, each with a different infrastructural prject completed

6.A.3 Other Coefficients

For each sector, we have supplied three coefficients exogenously to the model. These coefficients are given here.

The share of goods transport per sector π_s has been determined by outside experts using figures about transport costs from the available data on different sectors. They are given in the first column of table 6.7.

The tradeability of goods indicates the percentage of the output of a sector that can reasonably be expected to be available to customers outside the major region of production. For personal services like the proverbial haircut, tradeability is extremely low. Sector-wide figures are derived from IO date and sectorial indicators. It is given in the second column.

Finally, the degree to which a sector is exogenously tied to the present location is given in the third column. Exogenous ties result when the firm does not consider the price of inputs or its own price in the location decision. On the supply side, this happens when a firm uses specialized local inputs like natural resources or specialized labor, or a facility like a port. On the demand side, ties come about because of localized outputs, as for instance the local government sector can only supply its own citizens.

Sector	Share of goods transp. (π)	Tradeability	Exogenous ties
1 Agriculture	0.90	1.00	0.80
2 Mining	0.90	1.00	1.00
3 Industry	0.70	1.00	0.10
4 Public Utils	1.00	0.50	0.30
5 Construction	0.70	0.70	0.20
6 Trade & Repairs	0.30	0.50	0.30
7 Hotels etc.	1.00	0.50	0.50
8 Transport	0.70	0.75	0.30
9 Finance	0.00	0.70	0.30
10 Services	0.00	0.55	0.30
11 Government	0.10	0.45	1.00
12 Education	0.00	0.55	0.80
13 Health	0.10	0.45	0.80
14 Culture	0.10	0.55	0.50

Table 6.7: Exogenous coefficients for each sector

6.B Derivations

6.B.1 Costs and Production functions

When two factors are combined in the Cobb-Douglas production function $X^\alpha Y^{1-\alpha}$ and total costs $X \cdot p_X + Y \cdot p_Y$ are minimized, marginal costs are, up to a constant factor, equal to

$$p_X^\alpha \cdot p_Y^{1-\alpha}. \tag{6.14}$$

If you put the multiplicative constant in front of the production function, the marginal costs are exactly equal to (6.14). For this reason, c_{ps} and c'_{ps} are used. It is not hard to prove that they must be equal to

$$c_{ps} = (1-\alpha_{ps})^{\alpha_{ps}-1} \cdot (\alpha_{ps})^{-\alpha_{ps}}$$

and

$$c'_{ps} = \prod_{s'=1}^{14} \gamma_{p,s,s'}^{-\gamma_{p,s,s'}}.$$

6.B.2 The number of firms

Proof. We want to proof that the number of firms of a certain sector s in a community c in province p is proportional to the amount of labor consumed by that sector in that community multiplied by that sector's local labor requirement α_{ps} . Exit and entry are free so that each firm makes zero profits. It follows that each firm operates on a scale where gross profits $Y\cdot$ MC / $(\sigma_s - 1)$ are equal to a fixed startup cost F_s, which may differ per sector. From the equation for marginal costs, this implies that the optimal scale of firm j in sector s and region p, Y^*, is equal to

$$Y^*_{j,ps} = (\sigma_s - 1) \cdot F_s \cdot G_{c,s}^{a_s - 1} \tag{6.15}$$

From production function (6.3) we derive the production of a firm as a function of the amount of labor used. It turns out that this is

$$\begin{aligned} Y_{j,ps} &= c_{ps} \cdot L_j \cdot \left(\frac{w}{G_{c,s}}\right)^{1-\alpha_{ps}} \cdot \left(\frac{1-\alpha_{ps}}{\alpha_{ps}}\right)^{1-\alpha_{ps}} \\ &= \alpha_{ps} \cdot L_j \cdot G_{c,s}^{\alpha_{ps}-1} \end{aligned} \tag{6.16}$$

where we use $w = 1$ and the definition of c_{ps}. Equating (6.15) and (6.16) we find that a firm operating at optimal scale uses a fixed amount of labor, equal to

$$L^* = \frac{\sigma_s - 1}{\alpha_{ps}} F_s.$$

Because the amount of labor consumed in community c by sector s, $L_{c,s}^{total}$, is equal to

$$L_{c,s}^{total} = n_{cs} \cdot L^* \tag{6.17}$$

we find that the number of firms n_{cs} varies proportionally to the product of $L_{c,s}^{total}$ and α_{ps}. ■

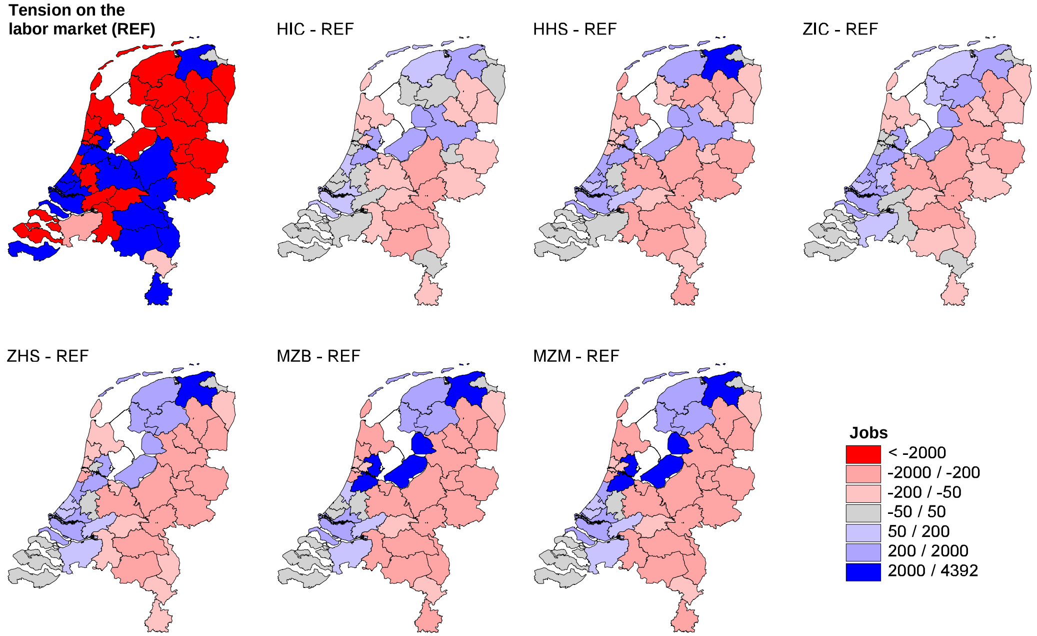

Figure 6.2: The effect of the six scenarios on the number of jobs in each COROP region.

Chapter 7

Summary and conclusions

7.1 Summary

Over time, two regularities may be observed in the world:

- *Economic activity is distributed unevenly on many scales.* There seems to be a force at work that pulls economic activity together, resulting in a system of cores and peripheries. Such a system may be observed at a worldwide scale, but also on the scale of countries and provinces.
- *On average, economies grow.* Technological advances, accumulation of capital, growing populations: for most of recorded history, world production has increased over time.

Chapter 2 illustrates these two trends with some data and then draws attention to the fact that, until recently, economic theory did not explain important aspects of either trend. Models of exogenous growth would describe the transition of economies to a steady-state growth path, but leave the determination of that path outside the model. In the end, productivity increases would come from technological improvements; their component in economic growth has been called 'a measure of our ignorance' by Abramovitz (1956).

Spatial models would describe the effect of a concentration on its *hinterland*, using the land rent gradient of von Thünen (1842) or the market potential function of Harris (1954). The occurrence of the concentration itself is explained by the Marshallian trinity of a shared labor market, the closeness of intermediate goods suppliers and pure external effects (Marshall 1920). The different sizes of these external effects, balanced against the congestion that discourages city growth, then explains the different sizes of cities (Henderson 1974). The assumption of pure, unobservable, external effects as the binding force in agglomerations, however, leaves the economist empty-handed when it comes to predicting the effect on cities of changes in transport costs or the occurrence of new industries.

In chapter 2 we explore two theories that provide new insights into these problems. Both theories make use of the monopolistic competition framework of Dixit and Stiglitz (1977), in which there are increasing returns on the level of the firm, as well as increasing returns to the number of firms (or, the number of varieties). These qualities of the model explain why there exist *complementarities* between firms, and these complementarities can be the source of agglomeration as well as growth. The two theories are the economic geography theory of Krugman (1991b) and endogenous growth theory of Romer (1986).

The beauty of the monopolistic competition model is that the owner of a single firm recognizes the fact that other firms complement his enterprise and uses this information in his decision, but no single firm can grow big enough to affect the size of the total 'complementarity'. Thus, it is a model where small actors, through their individual optimizations, bring about a macro-equilibrium with interesting qualities (Matsuyama 1993).

One such quality may be agglomeration; when there exist transport costs between different locations and individual firms value the complementarities of other firms, it is intuitive that there will exist an equilibrium in which all firms have moved to the same location. The channel through which these complementarities travel can differ between models: in paragraph 2.3.4 we identify the labor market, the market for intermediate goods and the R&D sector as possible media. The countervailing force to agglomeration in these models is local demand: even though being near other firms brings advantages, serving local demand from the peripheral region may be worthwhile. Another possible quality of the macro-equilibrium is lasting growth: when returns to variety imply that an increase in the number of firms makes all other firms more efficient, growth can be an equilibrium solution (paragraph 2.4).

Because both theories rest on the same foundation, prospects for a theory of growth *and* agglomeration are hopeful. In paragraph 2.5, several such theories are surveyed. The interplay between growth and location turns out to change the outcomes that are found in either literature by itself. Stable equilibria in static geography models turn out to be unstable in a dynamic context. On the other side, the rate of growth is influenced by the location pattern, which itself depends on initial values.

In chapter 3, we extend the economic geography model in which complementarities between firms exist on the market for intermediate goods. This type of model, introduced by Venables (1996b), explains agglomeration of economic activity when the population is not mobile between regions; this makes it suitable to investigate the differences in agglomeration between nations. In the original model, the simplifying assumption is made that firms use the output of *all* other firms as an intermediate input. We introduce an extension to the model, in which each firm can have its own input demand function, depending on its production process.

This extension introduces a wealth of possibilities into the model; there are infinitely many firms, and each of them can have a particular input demand function. Even with only two regions, it would be very hard to discuss every possible outcome of this model. We therefore proceed in two, simplifying, directions. For the remaining part of chapter 3, we adopt a number of simple, sweeping assumptions about the use of intermediate inputs by groups of firms. This allows us to asses the different ways in which differences in technology can affect the spatial equilibrium. We find that it is possible for two groups of firms to repel each other, when they do not use each other's output sufficiently. The externality that drives two unrelated sectors to different locations is their effect on local wages. In paragraph 3.3, we discuss how different technologies interact with location decisions if there is economic growth. Assuming a pattern of input demand that is consistent with technological progress, we show that in this model, firms locate according to their age: there exists a region in which firms use old technology and a region where new technology is introduced and used. When the number of firms increases, older firms in that region are to the *old* region.

The other direction in which we simplify the initial extension of chapter 3 is by letting go of the concept of 'continuous sectors' in chapter 4. If we assign each firm to one of a fixed number of sectors and stipulate that all firms in a sector demand the same intermediate inputs, we can contain the number of possible specifications. More precisely, under this assumption the demand for intermediate inputs can be summarized by an input-output (IO)-matrix of limited size. Thus when the number of sectors is N, the characteristics of intermediate-input demand are given by a parameter of dimension $N \times (N-1)$.[1] We present a full set of solutions in the case where $N = 2$ and there are two regions. Under those circumstances, there are four types of equilibrium (paragraph 4.3). We find that there exist a number of parameters in IO-space where the stability of these different equilibria switches. This means that a small change in technology can cause a dramatic change in the type of equilibrium that obtains. This disproportionate effect of a changing parameter has been shown before in the case of transport costs (Krugman 1991a). We show that it also applies to parameters in the production technology.

Next, we direct our attention away from the theoretical properties of geography models and turn to the estimation of their parameters in chapter 5. Paragraph 5.2 surveys earlier attempts to validate, estimate and parametrize these models in the literature. We discuss three approaches in detail: validation of the models by measuring the home market effect, estimation of parameters with the use of a spatial wage structure and the

---

[1] Since the columns of a matrix of input shares must sum to one, the dimension of this parameter is lower than the number of entries in the matrix, N^2.

art of parametrizing applied general equilibrium models.

In the rest of the chapter, we attempt to estimate a spatial wage structure with data on American states. We start by replicating the methodology of Redding and Venables (2001), who use data on world incomes and distance, in paragraph 5.5.1. Our dataset has the advantage of pertaining to regions of a single country, which eliminates error due to institutional differences or trade barriers between countries. The dataset is also smaller and we expect less variation between the observations. Our results confirm the conclusion of Redding and Venables that the model has reasonable explanatory power. However, we find that on the scale of US states this power is substantially less than when the model is applied to a sample of countries worldwide.

A problem with the methodology used to obtain these parameter estimates is the following: the model that you get when inserting these estimates does not necessarily agree with the original observation as a state of equilibrium. That is, if we are interested in finding a model that explains the data as good as possible, the parameter estimates obtained in this way are not necessarily the best. We introduce another methodology which might improve on this result in paragraph 5.5.2. In this paragraph, we search over a grid of parameters, computing the full general equilibrium solution for each parameter-candidate. The estimate is the parameter in which that solution deviates as little as possible from the data. Monte Carlo methods are used to generate standard errors for these estimates. We use a model, thus specified, to evaluate a number of counterfactuals and find the (spatial) effects of a localized change in wages and a drop in transport costs in paragraph 5.6.

With the last paragraph, we have entered into the field of policy evaluation. Now that we have formulated an understanding of geographical patterns in activity and found a way to estimate and calibrate a model to the present situation, we can run counterfactuals as a tool to evaluate the effects of policy. An exercise of this nature is the subject of chapter 6. In this chapter, we use the results of our previous work to specify a model of the Dutch economy with 548 regions and 14 sectors. Intermediate good demand for each sector is specified by an input-output matrix, which may differ between regions. We use this model to evaluate the construction of a railway between Amsterdam and Groningen in 2020. We find that, due to our neglect of the labor market in this model, we have to deal with an instability: left unchecked, the model agglomerates all activity in one location. We opt for a first-step approximation, in which we take some of the feedback out of the model.

7.2 Conclusions

In this thesis we look at economic geography models from a number of angles. We started by placing the theory in a context of preceding theories, both earlier work on spatial economics and other children of the monopolistic competition 'revolution.' Next, we looked at the theoretical properties of these models, especially when we allow firms to have different demand functions for intermediate goods. We estimated the model using a dataset on US states, and computed a number of counterfactuals. Finally, we used the theory to conduct a policy evaluation exercise concerning the construction of a railroad.

From the first chapter, we learn that there exist a large number of progenitors to the current crop of economic geography models. The models themselves are direct descendants of the earlier waves of MC-based innovation in the areas of industrial organization, international trade and economic growth. However, many of the concepts that these models formalize have been known and used for a long time by other theories, albeit in a less formal manner. Examples of some of these concepts are the gravity equation and the market potential function. An advantage of the formal microeconomic underpinning that the 'new' theory provides is that they may now be used in computations of consumer welfare, and embody explicit assumptions about economic behavior.

Expanding the model with varying types of intermediate demand, as we did in chapters 3 and 4, shows just how much the standard models depend on their simplifying assumptions. Relaxing one such simplifying assumption opens up a whole gallery of new models with different types of equilibria. Because solving these models often requires the use of numerical methods, it is complicated to map all possible outcomes after the change.

Nonetheless, we can draw some useful conclusions from the extensions in chapters 3 and 4. The models in chapter 3 show that it is possible to separate the agglomerating and dispersing forces that operate between firms. By specifying a particular input demand function, we can eliminate the attraction between two groups of firms and observe that they move to different locations. The dispersing force is the local wage rate; this is reminiscent of the practice of multinational enterprises to relocate their manufacturing to low-wage countries. The same model can be used to show that in a growing economy, it is possible that all innovative firms locate in the same region, leaving the other region with the older manufacturing processes. Which region gets the innovative firms is decided by history, and cannot easily be changed. This outcome can be used to argue against subsidies that would help 'backward' regions attract innovative firms in the hope of creating jobs. Unless these firms are relatively independent of other firms using related technology, these subsidies will have no lasting effect.

The models in chapter 4 illustrate that the spatial equilibrium in economic geography models indeed depends on the technology that is used in production, and the different demand functions for intermediate inputs. It goes further by showing that a gradual change in these functions does not, in general, change the equilibrium except at a few crucial values. That is, there exist situations in which a small change in the input-output parameters can have catastrophic consequences. This is not unlike the property of standard core-periphery models to be sensitive to the level of transport costs at particular break- and sustain points. We offer a map of the boundaries between different equilibria in IO-parameter space.

The empirical exercises in chapter 5 show that the economic geography model with intermediate goods is a reasonable description of the level of wages in, and the direction of trade between American states in 1997. We use two methods of estimation, one of which has been used before on a sample of countries worldwide (Redding and Venables 2001). We find that applying the same model to a (smaller) sample of US states leads to less conclusive results. In particular, the effect of the surrounding geography on one state's wages are hard to measure. This may be the effect of a dataset that is smaller and contains a few dominating regions.

We introduce a new estimation procedure that takes into account the general equilibrium properties of the model. Using this procedure, we find parameter values that indicate increased sensitivity to distance and the presence of a shared border, relative to the first estimation. The standard errors (computed using Monte Carlo methods) cast some doubt on the reliability of these estimates, however.

With the parametrized model of the United States, we run two counterfactuals involving changes in the (central) state of Illinois. We show that a fall in local wages sets off a chain of events, redistributing demand toward Illinois and its neighboring states, who enjoy cheaper inputs. For those neighbors however, the total result turns out to be negative as they also face a drop in demand from Illinois. Next, we simulate a fall in transport costs between the neighboring states of Illinois and Indiana. This benefits the affected states, who with their cheaper products help the surrounding states as well. But again these neighbors are worse off in the end, this time because national demand is shifted away from them, towards Illinois and Indiana. The different ways in which these changes in the economic environment impact the rest of the country are easily tracked and quantified with our model, showing its use a policy evaluation instrument.

Our exercise in chapter 6 shows the results of a policy evaluation for which a large-scale economic geography model was built. The model has a number of shortcomings, but did a reasonable job in tracking and quantifying the effects of six infrastructural projects. We found that the most ambitious plan would lead to a shift of about 8,000 jobs, gained at both ends of the new line. Furthermore, consumers in the North are better off because their access to services offered in the (more agglomerated) West has improved.

Bibliography

Abramovitz, M.: 1956, Resource and output trends in the United States since 1870, *American Economic Review Papers and Proceedings* **46**(2), 5–23.

Aghion, P. and Howitt, P.: 1992, A model of growth through creative destruction, *Econometrica* **60**(2), 323–51.

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

Alonso, W.: 1964, *Location and Land Use: Toward a General Theory of Land Rent*, Harvard University Press, Cambridge, MA.

Amiti, M.: 1997, Regional specialisation and technological leapfrogging, *mimeo*, La Trobe University, Victoria, Australia.

AND: 2000, *Reistijden en afstandenmatrix Nederland 1998*, AND Data Solutions, Rotterdam, NL. CD-ROM.

Anderson, J. E. and van Wincoop, E.: 2003, Gravity with Gravitas: A solution to the border puzzle, *American Economic Review* **93**(1), 170–192.

Anselin, L. and Florax, R. J.: 1995, Small sample properties of tests for spatial dependence in regression models: Some further results, *in* L. Anselin and R. J. Florax (eds), *New directions in Spatial Econometrics*, Springer Verlag, Berlin, GE, chapter 2, pp. 21–74.

Arrow, K. J. and Hahn, F. H.: 1971, *General Competitive Analysis*, Holden-Day, San Fransisco, CA.

Arrow, K. J. and Scitovsky, T.: 1969, *Readings in Welfare Economics*, George Allen and Unwin Ltd., London.

Baldwin, R. and Martin, P.: 2003, Agglomeration and regional growth, CEPR Working Paper # 3960, London, UK. Forthcoming in the *Handbook of Regional and Urban Economics: Cities and Geography*.

Baldwin, R. E. and Forslid, R.: 1997, The core-periphery model and endogenous growth, CEPR Working Paper # 1749, London, UK.

Barro, R. J. and Sala-i-Martin, X.: 1995, *Economic Growth*, McGraw-Hill, New York.

Blaug, M.: 1996, *Economic theory in retrospect*, fifth edn, Cambridge University Press, Cambridge, UK.

Brakman, S. and Garretsen, H.: 2003, Locatie en concurrentie: inleiding tot de preadviezen, *Preadviezen van de Koninklijke Vereniging voor de Staathuishoudkunde*, KVS, Amsterdam, pp. VII–XXXI.

Brakman, S. and Heijdra, B. J.: 2003, *The Monopolistic Competition Revolution in Retrospect*, Cambridge University Press, Cambridge, UK.

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

Brakman, S., Garretsen, H. and Schramm, M.: 2000, The empirical relevance of the new economic geography, testing for a spatial wage structure in Germany, *mimeo*, University of Nijmegen, University of Groningen, NL.

Brakman, S., Garretsen, H. and Schramm, M.: 2004, The strategic bombing of German cities during World War II and its impact on city growth, *Journal of Economic Geography* **4**(2), 1–19.

Brock, W. and Malliaris, A.: 1989, *Differential Equations, Stability and Chaos in Dynamic Economics*, North-Holland, Amsterdam.

Bröcker, J.: 1995, Chamberlinian spatial computable general equilibrium modelling: A theoretical framework, *Economic Systems Research* **7**(2), 137–150.

Bröcker, J.: 1998, How would an EU-membership of the Visegrád-countries affect Europe's economic geography?, *Annals of Regional Science* **32**, pp. 91–114.

Bröcker, J.: 1999, Trans-European effects of "trans-European networks": a spatial cge analysis, *mimeo*, Technical University of Dresden, Dresden, GE.

Buchanan, J. M. and Yoon, Y. J. (eds): 1994, *The Return to Increasing Returns*, The University of Michigan Press, Ann Arbor.

Cass, D.: 1965, Optimum growth in an aggregate model of capital accumulation, *Regional Economic Studies* **32**, 233–40.

CBS: 2002, Statistics Netherlands, statline database, Heerlen, NL.

Chamberlin, E. H.: 1933, *The Theory of Monopolistic Competition*, 1st edn, Harvard University Press, Cambridge, MA.

Chamberlin, E. H.: 1956, *The Theory of Monopolistic Competition*, 7th edn, Harvard University Press, Cambridge, MA.

Chipman, J. S.: 1970, External economies of scale and competitive equilibrium, *Quarterly Journal of Economics* **84**(3), 347–85.

Ciccone, A. and Hall, R. E.: 1996, Productivity and the density of economic activity, *American Economic Review* **86**(1), 54–70.

Cliff, A. and Ord, J.: 1973, *Spatial autocorrelation*, Pion Limited, London, UK.

Davis, D. R.: 1998, The home market, trade, and industrial structure, *American Economic Review* **88**(5), 1264–1276.

Davis, D. R. and Weinstein, D. E.: 1998a, Does economic geography matter for international specialization?, *mimeo*, School of Business Administration, University of Michigan, Ann Arbor, MI. A slightly different version is available as NBER Working Paper # 5706.

Davis, D. R. and Weinstein, D. E.: 1998b, Economic geography and regional production structure: An empirical investigation., Federal Reserve Bank of New York Staff Report # 40.

Davis, D. R. and Weinstein, D. E.: 1999, Economic geography and regional production structure: An empirical investigation, *European Economic Review* **43**(2), pp. 379–407.

Dean, R. D., Leahy, W. H. and McKee, D. L.: 1970, *Spatial Economic Theory*, The Free Press, New York, NY.

Deardorff, A. V.: 1995, Determinants of bilateral trade: Does gravity work in a neoclassical world?, NBER Working Paper # 5377, Cambridge, MA.

Dixit, A.: 2000, Some reflections on theories and applications of monopolistic competition, Paper presented at SOM conference *The Monopolistic Competition revolution after twenty-five years*, Groningen, NL.

Dixit, A. K. and Norman, V.: 1980, *Theory of International Trade*, Cambridge University Press, Cambridge, UK.

Dixit, A. K. and Stiglitz, J. E.: 1977, Monopolistic competition and optimum product diversity, *American Economic Review* **67**(3), 297–308.

Domar, E. D.: 1946, Capital expansion, rate of growth, and employment, *Econometrica* **14**, 137–47. Reprinted in Stiglitz and Uzawa (1969).

Duranton, G. and Puga, D.: 2003, Micro-foundations of urban agglomeration economies, Forthcoming in J.V. Henderson and J.-F. Thisse, *The Handbook of Regional and Urban Economics*, volume 4. North Holland, Amsterdam.

Efron, B. and Tibshirani, R. J.: 1993, *An Introduction to the Bootstrap*, Vol. 57 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, New York, NY.

Ellison, G. and Glaeser, E. L.: 1999, The geographic concentration of industry: Does natural advantage explain agglomeration?, *American Economic ReviewPapers and Proceedings* **89**(2), 311–316.

Englmann, F. C. and Walz, U.: 1995, Industrial centers and regional growth in the presence of local inputs, *Journal of Regional Science* **35**(1), 3–27.

ESPON: 2003, The role, specific situation and potentials of urban areas as nodes in a plycentric environment, Internet: `http://www.espon.lu/online/documentation/projects/thematic/1091/3-ir.1.1.1_part_1.pdf`.

Ethier, W. J.: 1982, National and international returns to scale in the modern theory of international trade, *American Economic Review* **72**(3), 389–405.

Fujita, M.: 1986, Urban land use theory, *in* J. Lesourne and H. Sonnenschein (eds), *Location Theory*, Harwood Academic Publishers, Chur, CH, pp. 73–149.

Fujita, M.: 1988, A monopolistic competition model of spatial agglomeration, *Regional Science and Urban Economics* **18**(1), 87–124.

Fujita, M. and Thisse, J.-F.: 1996, Economics of agglomeration, *Journal of the Japanese and International Economies* **10**, 339–78.

Fujita, M., Krugman, P. R. and Venables, A. J.: 1999, *The Spatial Economy: Cities, Regions and International Trade*, MIT Press, Cambridge, MA.

Gabszewicz, J. and Thisse, J.-F.: 1986, Spatial competition and the location of firms, *in* J. Lesourne and H. Sonnenschein (eds), *Location Theory*, Harwood Academic Publishers, Chur, CH., pp. 1–71.

Goldfeld, S. M.: 1984, Comment on *Modeling the Banking Firm, A Survey, Journal of Money, Credit and Banking* **16**(4), 609–11.

Green, H. J.: 1964, *Aggregation in Economic Analysis*, Princeton University Press, Princeton, NJ.

Greenhut, M. L.: 1952, Integrating the leading theories of plant location, *The Southern Economic Journal* **18**, 526–38. Reprinted in Dean et al. (1970).

Greenhut, M. L.: 1956, *Plant Location in Theory and Practice*, University of North Carolina Press, Chapel Hill, NC.

Grossman, G. M. and Helpman, E.: 1991, *Innovation and Growth in the Global Economy*, MIT Press, Cambridge, MA.

Hanson, G. H.: 1998, Market potential, increasing returns, and geographic concentration, NBER Working Paper # 6429, Cambridge, MA.

Hanson, G. H.: 1999, Market potential, increasing returns, and geographic concentration, *mimeo*, University of Michigan, Ann Arbor, MI.

Harris, C.: 1954, The market as a factor in the localization of industry in the United States, *Annals of the Association of American Geographers* **44**, 315–348.

Harrod, R.: 1939, An essay in dynamic theory, *The Economic Journal* **49**, 14–33. Reprinted in Stiglitz and Uzawa (1969).

Heijdra, B. J.: 1997, Toch een revolutie?, *Tijdschrift voor Politieke Ekonomie* **20**(1), 7–33.

Heijdra, B. J. and Yang, X.: 1993, Monopolistic competition and optimum product diversity: A comment, *American Economic Review* **83**(1), 295–301.

Helpman, E.: 1984, Increasing returns, imperfect markets, and trade theory, *in* R. W. Jones and P. B. Kenen (eds), *Handbook of International Economics*, North-Holland, Amsterdam, NL, chapter 7, pp. 325–65.

Helpman, E.: 1998, The size of regions, *in* D. Pines, E. Sadka and I. Zilcha (eds), *Topics in Public Economics: Theoretical and Applied Analysis*, Cambridge University Press, Cambridge, MA., chapter 2, pp. 33–54.

Helpman, E. and Krugman, P. R.: 1985, *Market Structure and Foreign Trade*, MIT Press, Cambridge, MA.

Henderson, J. V.: 1974, The sizes and types of cities, *American Economic Review* **64**, 640–656.

Hicks, J.: 1989, The assumption of constant returns to scale, *Cambridge Journal of Economics* **13**, 9–17.

Hirschman, A.: 1958, *The Strategy of Economic Development*, Yale University Press, New Haven, CT.

Hotelling, H.: 1929, Stability in competition, *The Economic Journal* **39**, 41–57. Reprinted in Dean et al. (1970).

Jacobs, J.: 1969, *The Economy of Cities*, Penguin Books, Harmondsworth, UK.

Jones, C. I.: 1995a, R & D-based models of economic growth, *Journal of Political Economy* **103**(4), 759–84.

Jones, C. I.: 1995b, Time series tests of endogenous growth models, *Quarterly Journal of Economics* **110**(2), 495–525.

Judd, K. L.: 1985, On the performance of patents, *Econometrica* **53**(3), 567–585.

Junius, K.: 1996, Economic development and industrial concentration: An inverted U-curve, Kiel Working Paper # 770, Kiel Institute of World Economics, Kiel, Germany.

Kaldor, N.: 1970, The case for regional policies, *Scottish Journal of Political Economy* **17**, 337–47.

Knaap, T.: 1997, Long term convergence in a long panel data set, Paper presented at the SOM workshop on Growth, Convergence and International Trade, Groningen University.

Koopmans, T. C.: 1965, On the concept of optimal economic growth, *(Study Week on the) Econometric Approach to Development Planning*, North-Holland Publishing Co., Amsterdam and Rand McNally, Chicago, chapter 4, pp. 225–87.

Krugman, P. R.: 1979, Increasing returns, monopolistic competition, and international trade, *Journal of International Economics* **9**(4), 469–479. Reprinted as Chapter 1 in Krugman (1990).

Krugman, P. R.: 1980, Scale economies, product differentiation, and the pattern of trade, *American Economic Review* **70**, 950–959. Reprinted as Chapter 2 in Krugman (1990).

Krugman, P. R.: 1990, *Rethinking International Trade*, The MIT Press, Cambridge, MA.

Krugman, P. R.: 1991a, *Geography and Trade*, Leuven University Press, Leuven, Belgium and MIT Press, Cambridge, MA.

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

Krugman, P. R.: 1995a, *Development, Geography, and Economic Theory*, The Ohlin Lectures, MIT Press, Cambridge, MA.

Krugman, P. R.: 1995b, Increasing returns, imperfect competition and the positive theory of international trade, *in* G. M. Grossman and K. Rogoff (eds), *Handbook of International Economics*, Vol. III, Elsevier Science, Amsterdam, chapter 24, pp. 1243–77.

Krugman, P. R.: 1998, Space: The final frontier, *Journal of Economic Perspectives* **12**(2), 161–74.

Krugman, P. R. and Venables, A.: 1996, Integration, specialization and adjustment, *European Economic Review* **40**, 959–967.

Krugman, P. R. and Venables, A. J.: 1995, Globalization and the inequality of nations, *Quarterly Journal of Economics* **110**(4), 857–80.

Lancaster, K.: 1979, *Variety, Equity and Efficiency*, Columbia University Press, New York, NY.

Lösch, A.: 1967, *The Economics of Location*, 2nd, rev. edn, John Wiley & Sons Science Editions, New York, NY. Translated from German by W. H. Woglom for Yale University Press.

Lucas, R. E.: 1988, On the mechanics of economic development, *Journal of Monetary Economics* **22**, 3–42.

Maddison, A.: 1982, *Phases of Capitalist Development*, Oxford University Press, Oxford (UK).

Maddison, A.: 1995, *Monitoring the World Economy, 1820-1992*, OECD, Paris.

Mankiw, N., Romer, D. and Weil, D. N.: 1992, A contribution to the empirics of economic growth, *Quarterly Journal of Economics* **107**, 407–37.

Marshall, A.: 1920, *Principles of Economics*, Macmillan, London.

Martin, P. and Ottaviano, G. I. P.: 1996a, Growing locations: Industry location in a model of endogenous growth, CEPR discussion paper # 1523, CEPR, London.

Martin, P. and Ottaviano, G. I. P.: 1996b, Growth and agglomeration, CEPR Working Paper # 1529, London, UK.

Matsuyama, K.: 1993, Modelling complementarity in monopolistic competition, *Bank of Japan Monetary and Economic Studies* **11**(1), 87–109.

Matsuyama, K.: 1995, Complementarities and cumulative processes in models of monopolistic competition, *Journal of Economic Literature* **33**, 701–29.

Migration News: 2001, EU: Freedom of movement, Internet: `http://migration.ucdavis.edu/mn/more.php?id=2387_0_4_0`.

Mion, G.: 2003, Spatial externalities and empirical analysis: The case of Italy, CORE Discussion Paper 2003/15, UCL Leuven. Internet: `http://www.core.ucl.ac.be/services/psfiles/dp03/dp2003-15.pdf`.

Mulligan, C. B. and Sala-i-Martin, X.: 1993, Transitional dynamics in two-sector models of endogenous growth, *Quarterly Journal of Economics* **108**(3), 739–73.

Myrdal, G.: 1957, *Economic Theory and Underdeveloped Regions*, Duckworth, London.

NASA Johnson Space Center: 2003, Image collection, The full size image is available from `http://eol.jsc.nasa.gov`, mission ISS006, roll E, frame 22939.

Neary, J. P.: 2001, Of hype and hyperbolas: Introducing the new economic geography, *Journal of Economic Literature* **39**(2), 536–561.

Neary, J. P.: 2003, Globalization and market structure, *Journal of the European Economic Association* **1**(2-3), 245–271.

Nonneman, W. and Vanhoudt, P.: 1996, A further augmentation of the Solow model and the empirics of economic growth for OECD countries., *Quarterly Journal of Economics* **111**(3), 943–53.

OECD: 2000, *International Trade by Commodities Statistics*, Vol. 3, OECD, Paris, France.

Oosterhaven, J., Elhorst, J. P., Knaap, T., Romp, W. E., Stelder, T. and Gerritsen, E.: 2000, Ruimtelijk economische effecten van zes zuiderzeelijn varianten, REG publication 20 , Stichting Ruimtelijke Economie, RuG, Groningen, NL.

Ottaviano, G. I. P. and Puga, D.: 1997, Agglomeration in the global economy: A survey of the 'new economic geography', Discussion Paper # 356, Centre for Economic Performance, London, UK.

Overman, H. G., Redding, S. and Venables, A. J.: 2001, The economic geograpy of trade, production and income: A survey of the empirics, Centre for Economic Performance, London School of Economics working paper.

Pack, H.: 1994, Endogenous growth theory: Intellectual appeal and empirical shortcomings, *Journal of Economic Perspectives* **8**(1), 55–72.

Puga, D.: 1999, The rise and fall of regional inequalities, *European Economic Review* **43**(2), 303–34.

Quah, D.: 2002, Spatial agglomeration dynamics, *American Economic Review Papers and Proceedings* **92**(2), 247–252.

Rebelo, S.: 1991, Long-run policy analysis and long-run growth, *Journal of Political Economy* **99**(3), 500–21.

Redding, S. and Schott, P. K.: 2003, Distance, skill deepening and development: Will peripheral countries ever get rich?, NBER Working Paper # 9447, Cambridge, MA.

Redding, S. and Venables, A.: 2001, Economic geography and international inequality, Centre for Economic Performance, London School of Economics discussion paper 495. Internet: `http://cep.lse.ac.uk/pubs/download/DP0495.pdf`.

Rensman, M.: 1996, Economic growth and technological change in the long run, SOM Research Report # 96C10, Groningen, The Netherlands.

Rivera-Batiz, F. L.: 1988, Increasing returns, monopolistic competition, and agglomeration economies in consumption and production, *Regional Science and Urban Economics* **18**(1), 125–53.

Rivera-Batiz, L. A. and Romer, P. M.: 1991, Economic integration and endogenous growth, *Quarterly Journal of Economics* **106**(2), 531–55.

Robinson, J.: 1933, *The economics of imperfect competition*, 1st edn, MacMillan, London, UK.

Romer, D.: 1993, The new keynesian synthesis, *Journal of Economic Perspectives* **7**(1), 5–22.

Romer, P. M.: 1986, Increasing returns and long-run growth, *Journal of Political Economy* **94**(5), 1002–37.

Romer, P. M.: 1987, Growth based on increasing returns due to specialization, *AEA Papers and Proceedings* **77**(2), 56–62.

Romer, P. M.: 1990, Endogenous technological change, *Journal of Political Economy* **98**(5, pt. 2), S71–102.

Romer, P. M.: 1991, Increasing returns and new developments in the theory of growth, *in* W. A. Barnett, B. Cornet, C. D'Aspremont, J. Gabszewicz and A. Mas-Colell (eds), *Equilibrium Theory and Applications: Proceedings of the Sixth International Symposium in Economic Theory and Econometrics*, Cambridge University Press, pp. 83–110.

Romer, P. M.: 1994, The origins of endogenous growth, *Journal of Economic Perspectives* **8**(1), 3–22.

RuG/CBS: 1999, *Regionale Samenhang in Nederland, Bi-Regionale Input-Output Tabellen En Aanbod- En Gebruiktabellen Voor de 12 Provincies En de Twee Mainport Regio's*, Rijksuniversiteit Groningen/Centraal Bureau voor de Statistiek, REG-publicatie 20, Stichting Ruimtelijke Economie Groningen, Groningen, NL and Heerlen, NL.

Samuelson, P. A.: 1952, The transfer problem and transport costs: The terms of trade when impediments are absent, *The Economic Journal* **62**, 278–304.

Samuelson, P. A.: 2001, A Ricardo-Sraffa paradigm comparing gains from trade in inputs and finished goods, *Journal of Economic Literature* **39**, 1204–1214.

Schumpeter, J. A.: 1942, *Capitalism, Socialism and Democracy*, Harper & bros., New York, NY.

Scitovsky, T.: 1954, Two concepts of external economies, *Journal of Political Economy* **62**, 143–51.

Smith, A.: 1776, *An Inquiry into the Nature and Causes of the Wealth of Nations*, Clarendon Press, Oxford.

Smith, S. D.: 1997, *Statistical summary, Minerals Yearbook (II)*, US Geological Survey, Dept. of the Interior, Reston, VA. internet: `http://minerals.er.usgs.gov/minerals/pubs/commodity/statistical_summary/871497.pdf`.

Solow, R. M.: 1956, A contribution to the theory of economic growth, *Quarterly Journal of Economics* **70**, 65–94. Reprinted in Stiglitz and Uzawa (1969).

Solow, R. M.: 1994, Perspectives on growth theory, *Journal of Economic Perspectives* **8**(1), 45–54.

Spence, M.: 1976, Product selection, fixed costs, and monopolistic competition, *Regional Economic Studies* **43**, 217–35.

Sraffa, P.: 1960, *Production of Commodities by Means of Commodities, Prelude to a Critique of Economic Theory*, Cambridge University Press, Cambridge, MA.

Starrett, D.: 1978, Market allocation of location choice in a model with free mobility, *Journal of Economic Theory* **17**, 21–37.

Stigler, G.: 1968, *The Organization of Industry*, The University of Chicago Press, Chicago, IL and London, UK.

Stigler, G. J.: 1951, The division of labor is limited by the extent of the market, *Journal of Political Economy* **59**(3), 185–93. Reprinted in Buchanan and Yoon (1994).

Stiglitz, J. E. and Uzawa, H.: 1969, *Readings in the Modern Theory of Economic Growth*, The MIT Press, Cambridge, MA and London, UK.

Summers, R. and Heston, A.: 1988, A new set of international comparisons of real product and price levels estimates for 130 countries, *Review of Income and Wealth* **34**, 1–26.

Swan, T. W.: 1956, Economic growth and capital accumulation, *The Economic Record* **32**(63), 334–61. Reprinted in Stiglitz and Uzawa (1969).

Takayama, A.: 1985, *Mathematical Economics*, 2nd edn, Cambridge University Press, New York, NY.

Tinbergen, J.: 1962, *Shaping the World Economy: Suggestions for an International Economic Policy*, The Twentieth Century Fund, New York, NY.

van Oort, F. G.: 2002, Agglomeration, economic growth and innovation. spatial analysis of growth- and R&D externalities in the Netherlands, Dissertation, Tinbergen Institute, Erasmus University Rotterdam.

Varian, H. R.: 2000, *Variants in Economic Theory*, Edward Elgar Publishing, Cheltenham, UK.

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

Venables, A. J.: 1996b, Trade policy, cumulative causation, and industrial development, *Journal of Development Economics* **49**, 179–197.

Venables, A. J.: 2000, Cities and trade: External trade and internal geography in developing economies, Background paper written for 1999 World Development Report, 'Entering the 21st century'. *mimeo*, London School of Economics.

Venables, A. J. and Gasiorek, M.: 1996, Evaluating regional infrastructure: A computable equilibrium approach, *mimeo*, London School of Economics, London, UK.

von Thünen, J. H.: 1842, *Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*, Gustav Fischer Verlag, Jena.

Weber, A.: 1909, *Über den Standort der Industrien*, Mohr Verlag, Tübingen.

Weitzman, M. L.: 1994, Monopolistic competition with endogenous specialization, *Review of Economic Studies* **61**, 45–56.

Wolf, H. C.: 1997, Patterns of intra- and inter-state trade, NBER Working Paper # 5939, Cambridge, MA.

World Bank: 2003, World development indicators, Washington, DC, `http://www.worldbank.org`.

Young, A.: 1993, Substitution and complementarity in endogenous innovation, *Quarterly Journal of Economics* **108**(3), 775–807.

Young, A. A.: 1928, Increasing returns and economic progress, *The Economic Journal* **38**, 527–42. Reprinted in Arrow and Scitovsky (1969).

Samenvatting

Een van de opmerkelijkste kenmerken van economische activiteit is de ongelijkmatige ruimtelijke verdeling. In tabel 1.1 op pagina 2 staan gegevens over de productie per oppervlakte van verschillende continenten, landen en provincies. Op ieder schaalniveau is de verdeling ongelijk, waarbij de minst productieve eenheid altijd minder dan 10% van de meest productieve eenheid haalt.

Een tweede lange-termijn kenmerk van vrijwel elke economie is een bestendige groei. In hoofdstuk 2 van dit proefschrift wordt met behulp van empirische data ingegaan op de beide trends. We beschouwen bovendien de achterliggende economische theorie van agglomeratie en economische groei. In alletwee de gevallen is er recent een verandering opgetreden in de manier waarop vanuit de theorie naar deze verschijnselen wordt gekeken. Voor wat betreft economische groei is ten behoeve van analyse lang gebruik gemaakt van het raamwerk van Solow (1956), waarin groei op de lange termijn veroorzaakt wordt door technologische vooruitgang. De bron van deze vooruitgang ligt buiten het model; deze theorie staat daarom bekend als *exogene* groeitheorie. Sinds het werk van Romer (1986) is er een alternatief, de *endogene* groeitheorie, dat de oorzaken van groei binnen het model vervat.

De theorie van de locatie van economische activiteit is recentelijk ook op een dergelijke manier uitgebreid. De modellen van von Thünen (1842) en Harris (1954) beschrijven de effecten die een agglomeratie heeft op de ruimte er omheen, maar gaan niet in op de oorzaken die ervoor hebben gezorgd dat de agglomeratie in de eerste plaats is ontstaan. Er bestaan wel ideeën omtrent de oorzaken van ruimtelijke verschillen in economische activiteit, onder meer die van Marshall (1920), maar deze ontbeerden tot voor kort een modelmatige onderbouwing. De zogenoemde 'Marshalliaanse drie-eenheid' van redenen voor agglomeratie bestaat uit:

- Externe effecten. Externe effecten zijn effecten van het ene bedrijf op het andere die niet in de overweging van het eerste bedrijf worden betrokken. Voorbeelden zijn het bijdragen tot een 'innovatieve atmosfeer' of een bijdrage aan de bekendheid van een regio door een enkel bedrijf, tot voordeel van alle andere bedrijven.

- Een gespecialiseerde arbeidsmarkt. Een cluster van bedrijven uit een bepaalde sector die zich dicht bij elkaar vestigen heeft als voordeel dat arbeid, gespecialiseerd in het werk van deze sector, een prikkel heeft zich ook in deze omgeving op te houden. Op die manier ontstaat een liquide arbeidsmarkt, waarvan beide partijen voordeel hebben.

- Intermediaire leveringen. Als alle bedrijven in een bepaalde sector een bepaald intermediair goed nodig hebben voor hun productie, dan is het voor de leverancier van dit goed prettig als bedrijven uit de sector zich dicht bij elkaar vestigen. Gegeven de agglomeratie van afnemers ontstaat er, op dezelfde locatie, een agglomeratie van leveranciers. Geen van de partijen ondervindt dan een prikkel om zich ergens anders te vestigen.

Een evenwicht tussen (agglomeratie-veroorzakende) externe effecten en de congestie die gepaard gaat met een agglomeratie kan worden gebruikt om het bestaan van steden van verschillende omvang te verklaren (Henderson 1974). Het probleem met 'pure' externe effecten is echter dat ze in de praktijk moeilijk meetbaar zijn, waardoor empirische verificatie van het model uitblijft.

In alle onderdelen van de drie-eenheid is er sprake van *complementariteit* tussen de bedrijven. De economische theorie van complementariteit (zie Matsuyama 1993) laat zien dat er in dit geval meerdere economische evenwichten mogelijk zijn. Het consistent modelleren van complementariteit is echter pas recent goed mogelijk geworden door de theorie van monopolistische concurrentie van Dixit en Stiglitz (1977). Tot die tijd bleek het vervatten van een onderdeel van de bovenstaande drie-eenheid in een consistent economisch model te ingewikkeld.

Een geslaagde poging om het mechanisme van Marshall te formaliseren is Krugman (1991a). Dit is overigens niet de eerste in zijn soort, zie bijvoorbeeld Duranton en Puga (2003). De auteur introduceert de naam 'Economische Geografie' voor de door hem geïntroduceerde klasse van modellen (zie hierover ook Brakman and Garretsen 2003). Met gebruik van de theorie van Dixit en Stiglitz komt Krugman tot een model waarin, door de interactie tussen schaalvoordelen en transportkosten, een agglomeratie kan ontstaan. In welke regio deze agglomeratie ontstaat ligt niet van tevoren vast: de uitkomst van het model is pad-afhankelijk. Als de agglomeratie zich eenmaal heeft gevormd, is zij echter bestendig.

De oorzaak van de agglomeratie is in deze theorie in feite de complementariteit tussen bedrijven. Hoofdstuk 2 gaat in op de verschillende kanalen waardoor die complementariteit zich kan uiten: de arbeidsmarkt en de markt voor intermediaire goederen (zoals beschreven door Marshall) maar ook door de sector Onderzoek en Ontwikkeling.

Omdat ook endogene groeimodellen gebruikmaken van het model van Dixit en Stiglitz, lijkt het logisch om een theorie te ontwikkelen waarin de interactie tussen agglomeratie en economische groei wordt onderzocht. Een overzicht van de literatuur over dit soort theorieën wordt gegeven in paragraaf 2.5.

Hoofdstuk 2 toont onder meer aan dat veel resultaten uit de economische geografie al langer bekend waren, maar nog niet eerder in een model geformaliseerd. De zwaartekracht-vergelijking en het begrip 'marktpotentiaal', begrippen uit de internationale en regionale economie, zijn bijvoorbeeld ook af te leiden uit de economische geografie. Een voordeel van de geformaliseerde concepten is dat de modellen een microeconomische basis hebben, waardoor het mogelijk wordt om uitspraken te doen over welvaart.

Hoofdstuk 3 brengt een verfijning aan in het model van Venables (1996b), dat agglomeratie op basis van intermediaire goederen beschrijft. In het originele model wordt de aanname gemaakt dat alle bedrijven eenzelfde pakket intermediaire goederen en diensten gebruiken bij hun productie. Door toe te laten dat verschillende bedrijven hun eigen behoeften hebben, ontstaat een nieuw model waarvan de resultaten wezenlijk anders zijn dan in het model van Venables. Van enkele simpele voorbeelden wordt aangetoond dat, afhankelijk van de parameters, verschillende evenwichten bestaan. Het kan zijn dat de sectoren agglomereren in een eigen regio, maar het is ook mogelijk dat er een gemengde economie ontstaat waarin alle sectoren in elke regio aanwezig zijn. Hoewel deze evenwichten zeer verschillend zijn, liggen ze in het model dicht bij elkaar: een kleine verandering van de parameters kan resulteren in een dramatische verandering van het ruimtelijk evenwicht.

Deze modellen laten zien wat de verspreidende krachten in een economie kunnen zijn: wanneer de ruimtelijke verdeling van werknemers gegeven is, is de arbeidsmarkt bijvoorbeeld een verspreidend element: hogere lonen maken een agglomeratie minder aantrekkelijk als vestigingsplek.

In paragraaf 3.3 gaan we in op de vraag wat de introductie van verschillende sectoren in het model voor een gevolgen heeft als er in het model ook economische groei plaatsvindt. Daarvoor wordt een klein model met een specifieke samenhang tussen de sectoren ingezet. Deze samenhang schrijft voor dat bedrijven meer met elkaar verbonden zijn als ze uit hetzelfde tijdperk stammen. In dit model is het mogelijk om te laten zien dat er twee soorten regio's bestaan: één met 'oude' bedrijven en één met nieuwe bedrijven. Als er, door de groei, meer nieuwe bedrijven bijkomen in die laatste regio, verhuizen de oudste aanwezige bedrijven daar vandaan. Op die manier blijft het evenwicht bestaan.

Dit resultaat kan dienen als argument tegen subsidies die technologisch geavanceerde bedrijven naar een achtergebleven gebied moeten lokken, in de hoop dat dit de banengroei stimuleert. Het model toont aan dat er waarschijnlijk geen blijvende effecten van een eenmalige subsidie zijn: tenzij de aangetrokken bedrijven volledig onafhankelijk zijn van andere sectoren, zullen ze na verloop van tijd vertrekken.

Het volgende hoofdstuk borduurt voort op het model met meerdere typen bedrijven. In hoofdstuk 4 worden de bedrijven toegewezen aan een beperkt aantal sectoren, waartussen handel in intermediaire goederen bestaat. De samenstelling van die handel wordt beschreven in een zogenaamde input-output matrix. Ook in dit model geldt dat er verschillende uitkomsten mogelijk zijn, afhankelijk van de parameters. In een model met twee sectoren onderscheiden we vier typen evenwicht (zie de figuur op pagina 102). We brengen, voor alle mogelijke IO-matrices, in kaart welke van de evenwichten stabiel is. Het vaststellen van stabiliteit is een numerieke procedure, waarin de eigenwaarden van een matrix van afgeleiden bepalend zijn.

Het hoofdstuk toont aan dat de extreme gevoeligheid voor parameters, die modellen uit de economische geografie kenmerkt, zich uitstrekt tot de input-output matrix. Het brengt de stabiele evenwichten voor alle mogelijke matrices in beeld.

Na deze studie van de theoretische eigenschappen van modellen van economische geografie, stappen we over op een analyse van het empirische belang van de resultaten. Paragraaf 5.2 geeft een overzicht van de empirische literatuur op het gebied van economische geografie. Er zijn drie methoden te onderscheiden: studies die proberen het zogenaamde *home market effect* aan te tonen, dat een gevolg is van monopolistische mededinging; schattingen van parameters op basis van de ruimtelijke verdeling van de loonvoet; en het reproduceren van geobserveerde situaties door middel van de kalibratie van een algemeen evenwichtsmodel.

In hoofdstuk 5 proberen we verder een verklaring te vinden voor de ruimtelijke verdeling van lonen in de Verenigde Staten op basis van een model uit de economische geografie. Daarvoor worden twee econometrische methoden gebruikt. Ten eerste voeren we de schatting die Redding en Venables (2001) voor een wereld-dataset deden opnieuw uit voor de Verenigde Staten. Hieruit blijkt dat de kracht van het achterliggende model om de observaties te verklaren redelijk is, maar minder groot dan in de oorspronkelijke studie. Daarna wordt een nieuwe schattingsmethode gebruikt, die uitgaat van een consistente modeloplossing. De parameters worden geschat door de afstand tussen de modeloplossing en de data te minimaliseren. De standaardfout van de schatting wordt berekend met een numerieke 'Monte Carlo'-methode. Met de schattingsresultaten voor de parameters van het model voeren we een aantal simulaties uit. We simuleren de ruimtelijke effecten van verandering van het loon in één, centrale, staat. Verder bestuderen we de gevolgen van een verlaging van de transportkosten tussen twee staten.

De gematigde resultaten met de methode van Redding en Venables kunnen te wijten zijn aan het feit dat de gebruikte dataset (de Verenigde Staten) een kleiner aantal waarnemingen bevat dan het origineel (alle landen van de wereld). Ook kan het zijn dat een aantal waarnemingen van grote staten de schatting domineert. De methode waarbij het modelevenwicht wordt ingezet is nog niet eerder gebruikt. Bij beide methoden is de conclusie dat het gebruikte model een redelijke beschrijving van de ruimtelijke verdeling van de Amerikaanse lonen geeft. De simulaties laten bovendien zien hoe de effecten van een verandering in die loonstructuur zich over het land verspreiden.

In hoofdstuk 6 bespreken we de resultaten van een vergelijkbare empirische exercitie. In dat hoofdstuk wordt een model van Nederland opgezet, dat wordt gekalibreerd op een (veronderstelde) waarneming uit het jaar 2020. We gebruiken dat model vervolgens om de effecten van een nieuwe (zweef-)treinverbinding tussen de Randstad en het noorden van Nederland te onderzoeken. Het model in hoofdstuk 6 is op bepaalde vlakken zeer uitgebreid: we simuleren alle 548 Nederlandse gemeenten en maken onderscheid tussen veertien verschillende sectoren. Iedere provincie heeft een eigen input-outputtabel, waarmee het model een uitgebreide toepassing van de methoden van hoofdstuk 4 wordt. Verder maken we onderscheid tussen twee soorten transport, goederen en personen, waarbij elke sector een eigen combinatie van deze twee transportbehoeften gebruikt.

Een nadeel van het model is de afwezigheid van een corrigerend loonmechanisme op de arbeidsmarkt, waardoor er geen tegenkracht tegen de agglomeratie is. We simuleren daarom alleen de eerste stap, die het gevolg is van de constructie van een nieuwe verbinding. Het meest ambitieuze plan leidt in deze simulatie tot een verplaatsing van 8000 banen naar Noord-Nederland en de Randstad.