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.