1

Agglomeration and Economic Theory

1.1 INTRODUCTION

Just as matter in the solar system is concentrated in a small number of bodies (the planets and their satellites), economic life is concentrated in a fairly limited number of human settlements (cities and clusters). Furthermore, paralleling large and small planets, there are large and small settlements with very different combinations of firms and households. This book is a study of the reasons for the existence of a large variety of economic agglomerations. Even though economic activities are, to some extent, spatially concentrated because of natural features (think of rivers and harbors), our goal is to focus on economic mechanisms yielding agglomeration by relying on the trade-off between various forms of increasing returns and different types of mobility costs.

One should keep in mind that the concept of economic agglomeration refers to very distinct real-world situations.[1] At one extreme lies the core–periphery structure corresponding to North–South dualism. For example, Hall and Jones (1999) observed that high-income nations are clustered in small industrial cores in the Northern Hemisphere and that productivity per capita steadily declines with distance from these cores.

As noted by many historians and development theorists, economic growth tends to be localized. This is especially well illustrated by the rapid growth of East Asia during the last few decades. We view East Asia here as comprising Japan and nine other countries, that is, Republic of South Korea, Taiwan, Hong Kong, Singapore, Philippines, Thailand, Malaysia, Indonesia, and China. In 1990, the total population of East Asia was about 1.6 billion. With only 3.5% of the total area and 7.9% of the total population, Japan accounted for 72% of the gross domestic product (GDP) and 67% of the manufacturing GDP of East Asia. In Japan itself, the economy is very much dominated by its core regions formed by the five prefectures containing the three major metropolitan areas of Japan: Tokyo and Kanagawa prefectures, Aichi prefecture (containing the Nagoya metropolitan area), and Osaka and Hyogo prefectures. These regions

account for only 5.2% of the area of Japan but for 33% of its population, 40% of its GDP, and 31% of its manufacturing employment. Hence, for the whole of East Asia, the Japanese core regions with a mere 0.18% of the total area accounted for 29% of East Asia's GDP.

Strong regional disparities within the same country imply the existence of agglomerations at another spatial scale. For example, in Korea, the capital region (Seoul and Kyungki Province), which has an area corresponding to 11.8% of the country and includes 45.3% of the population, produces 46.2% of the GDP. In France, the contrast is even greater: the Île-de-France (the metropolitan area of Paris), which accounts for 2.2% of the area of the country and 18.9% of its population, produces 30% of its GDP. Inside the Île-de-France, only 12% of the available land is used for housing, plants, and roads, the remaining land being devoted to agriculture, forestry, or natural activities.

Regional agglomeration is also reflected in large varieties of cities, as shown by the stability of the urban hierarchy within most countries (J. Eaton and Eckstein 1997; Dobkins and Ioannides 2000). Cities themselves may be specialized in a very small number of industries, as are many medium-size American cities (Henderson 1997a). However, large metropolises like New York or Tokyo are highly diversified in that they nest many industries that are not related through direct linkages (Chinitz 1961; Fujita and Tabuchi 1997). Industrial districts involving firms with strong technological, or informational linkages, or both (e.g., the Silicon Valley or Italian districts engaged in more traditional activities) as well as factory towns (e.g., Toyota City or IBM in Armonk, New York) manifest various types of local specialization. Therefore, it appears that highly diverse size and activity arrangements exist at the regional and urban levels.

At a very detailed extreme of the spectrum, agglomeration arises under the form of large commercial districts set up in the inner city itself (think of Soho in London, Montparnasse in Paris, or Ginza in Tokyo). At the lowest level, restaurants, movie theaters, or shops selling similar products are clustered within the same neighborhood, not to say on the same street, or the clustering may take the form of a large shopping mall. Understanding such phenomena is critical for the design of effective urban policies.

The economic reasons that stand behind such strong geographical concentrations of consumption and production are precisely what we aim to investigate in this book. To achieve this objective, we will appeal to the concepts and tools of modern microeconomics. Because clusters appear at different geographical scales and involve various degrees of sectoral details, it would be futile to look for *the* model explaining different types of economic agglomerations (Papageorgiou 1983). This should not come as a surprise, for geographers have long known that geographical scale matters.[2] What is true at a certain spatial scale is not necessarily true at another (the "ecological fallacy"). For example, whether Los Angeles or Chicago may be considered as a megacenter or as a

collection of several large subcenters depends very much on the scale of observation. Likewise, during the 1980s the income differentials have decreased across country members of the European Union but not across regions within countries. The reason for such differences probably lies in the nature and balance of the system of forces at work at a given level of analysis. Or, in the words of Anas, Arnott, and Small (1998, 1440):

> It may be that the patterns that occur at different distance scales are influenced by different types of agglomeration economies, each based on interaction mechanisms with particular requirements for spatial proximity.

Yet, as will be seen, a few general principles seem to govern the formation of distinct agglomerations even though the content and intensity of the forces at work may vary with place and time.[3]

1.2 CITIES: PAST AND FUTURE

Casual observation reveals the extreme variation in the intensity of human settlements and land use – a fact that has culminated in the existence of *cities* in which population densities are very high.[4]

From a historical perspective, cities emerged in several parts of the world about 7,000 years ago as the consequence of the rise in agricultural surplus. The mere existence of cities may be viewed as a universal phenomenon whose importance slowly but steadily increased during the centuries preceding the sudden urban growth that appeared during the nineteenth century in a small corner of Europe (Bairoch 1985, chaps. 15–17). Technological development was necessary to generate the agricultural surplus without which cities would have been inconceivable at the time, as they would be today.

In addition to technological innovations, a fundamental change in social structure was also necessary: the division of labor into specialized activities. In this respect, there seems to be a large agreement among economists, geographers, and historians to consider "increasing returns" as the most critical factor in the emergence of cities. For example, J. Marshall (1989, 25) has suggested that

> quite apart from considerations related to defense, to royal whim, or to the supposed sacred importance of certain sites, the formation of towns made good economic sense in promoting a level of efficiency in commerce, manufacturing, and administration that would have been impossible to achieve with a completely dispersed population.

Although the sources are dispersed, not always trustworthy, and hardly comparable, data clearly converge to show the existence of an urban revolution. In Europe, the proportion of the population living in cities increased very slowly from 10% in 1300 to 12% in 1800 (Bairoch 1985). It was approximately 20% in 1850, 38% in 1900, 52% in 1950, and is now close to 75%, thus showing an explosive growth in the urban population (Bairoch 1985; United Nations 1994). In the United States, the rate of urbanization increased from 5% in 1800 to more

than 60% in 1950 and is now nearly 77%. In Japan, the rate of urbanization was about 15% in 1800 (Bairoch 1985), 50% in 1950, and is now about 78% (United Nations 1994). The proportion of the urban population in the world increased from 30% in 1950 to 45% in 1995 and will exceed 50% in 2005 (United Nations 1994). The world's urban population increases each year by the equivalent of 40 million (i.e., the population of Spain).

Furthermore, concentration in very big cities keeps rising. In 1950, only two cities had populations greater than 10 million: New York and Greater London. In 1995, fifteen cities belonged to this category. The largest one, Tokyo, with more than 26 million, exceeds the second one, New York, by 10 million. In 2025, 26 megacities will exceed 10 million in population (United Nations 1994).

Economists and geographers must explain why firms and households concentrate in large metropolitan areas even though empirical evidence suggests that the cost of living in such areas is typically higher than in smaller urban areas (Richardson 1987). As Lucas (1988, 39) neatly put it, "What can people be paying Manhattan or downtown Chicago rents for, if not for being near other people?" But Lucas did not explain why people want, or need, to be near other people. Likewise, economists and geographers must explain the formation of small and specialized clusters of firms and workers not necessarily located within major cities – such as many of the Italian industrial districts (Pyke, Becattini, and Sengenberger 1990, chap. 3) – and that appear to be very efficient in terms of productivity.

The increasing availability of high-speed transportation infrastructure and the fast-growing development of new informational technologies might suggest that our economies are entering an age that will culminate in the "death of distance." If so, locational difference would gradually fade because agglomeration forces would be vanishing. In other words, cities would become a thing of the past. We will see in this book that things are not that simple because the opposite trend may just as well arise. Indeed, one of the general principles to be derived from our analysis is that the relationship between the decrease in transport costs and the degree of agglomeration of economic activities is not that expected by many analysts: *Agglomeration happens provided that transport costs are below some critical threshold*,[5] although further decreases may yield dispersion of some activities owing to factor price differentials. In addition, technological progress brings about new types of innovative activities that benefit most from being agglomerated and, therefore, tend to arise in developed areas. Consequently, the wealth or poverty of nations seems to be more and more related to the development of prosperous and competitive clusters of specific industries as well as to the existence of large and diversified metropolitan areas (Glaeser 1998; Porter 1998, chaps. 6 and 7; Thisse and van Ypersele 1999).

The recent attitude taken by several institutional bodies and medias seems to support this view. For example, in its recent *World Development Report*, the World Bank (2000) stressed the importance of economic agglomerations and

cities for boosting growth and escaping from the poverty trap. Another example of this increasing awareness of the relevance of cities in modern economies can be found in *The Economist* (1995, 18):

> The liberalization of world trade and the influence of regional trading groups such as NAFTA and the EU will not only reduce the powers of national governments, but also increase those of cities. This is because an open trading system will have the effect of making national economies converge, thus evening out the competitive advantage of countries, while leaving those of cities largely untouched. So in the future, the arenas in which companies will compete may be cities rather than countries.

In this book, we intend to address the main causes for the formation of the various types of economic agglomerations described above. As discussed in the next two sections, this includes increasing returns to scale, externalities, and imperfectly competitive markets with general and strategic interdependencies. From this list, it should be clear that the economics of agglomeration is fraught with most of the difficulties encountered in economic theory.

Moreover, as will be seen in various chapters of this book, models of agglomeration involve both *complementarity* and *substitution* effects. For a long time, economists had problems handling complementarity effects, which can hardly be taken in account in the general competitive framework. This observation will lead us, in Section 1.4, to survey the rather complex history of the relationship between space and economic theory. Although space has not been ignored by some prominent economists, it has seldom been mentioned in economics texts. Thus, it is interesting to determine why this important ingredient of social life has been put aside for so long.

1.3 WHY DO WE OBSERVE AGGLOMERATIONS?

Intuitively, it should be clear that the spatial configuration of economic activities is the outcome of a process involving two opposing types of forces, that is, *agglomeration* (or centripetal) forces and *dispersion* (or centrifugal) forces. The observed spatial configuration of economic activities is then the result of a complicated balance of forces that push and pull consumers and firms. This view agrees with very early work in economic geography. For example, in his *Principes de géographie humaine* published posthumously in 1921, the famous French geographer Vidal de la Blache argued that all societies, rudimentary or developed, face the same dilemma: Individuals must get together to benefit from the advantages of the division of labor, but various difficulties restrict the gathering of many individuals.

1.3.1 Agglomeration and Increasing Returns

One would expect trade theory to be the branch of economics that has paid most attention to the spatial dimension. The reason is that changes in the conditions under which commodities are shipped, as well as changes in the mobility of

factors, affect the location of industry, the geography of demand and, eventually, the pattern of trade. The opposite has been true, for neoclassical trade theory has treated each country as dimensionless and has given little attention to the impact of trade costs. Yet, some predominant contributors in the field have long argued that location and trade are closely related topics. For example, Ohlin (1933; 1968, 97) has challenged the common wisdom that considers international trade theory as separate from location theory:[6]

> International trade theory cannot be understood except in relation to and as part of the general location theory, to which the lack of mobility of goods and factors has equal relevance.

Natural resources, and more generally production factors, are not uniformly distributed across locations, and it is on this unevenness that most of trade theory has been built.[7] The standard model of trade considers a setting formed by two countries producing two goods by means of two factors (labor and capital) under identical technologies subject to constant returns to scale and strictly diminishing marginal products. When factors are spatially immobile and goods can be costlessly moved from one country to the other, this model predicts the equalization of factor prices when the ratios of factor endowments are not too different.

Similarly, regional economics has long been dominated by the dual version of the neoclassical trade model. It is assumed that a single good is produced and that (at least) one production factor can *freely* move between regions. According to this model, capital flows from regions where it is abundant to regions where it is scarce until capital rents are the same across regions, or regional wage differences push and pull workers until the equalization of wages between regions is reached. Because the production function is linear homogeneous and has strictly diminishing marginal product in each factor, the marginal productivity of the mobile factor depends only on the capital–labor ratio. This implies that the mobile factor moves from regions with low returns toward regions with high returns up to the point at which the capital–labor ratio is equalized across all regions. In other words, the perfect mobility of one factor would be sufficient to guarantee the equalization of wages and capital rents in the interregional marketplace.[8]

Thus, it would seem that either costless trade or the perfect mobility of one factor would be sufficient to guarantee the convergence of labor income across various places.[9] Ignoring unevenness in the spatial distribution of natural resources, Mills (1972a, 4) very suggestively described this strange "world without cities" that would characterize an economy operating under constant returns and perfect competition as follows:

> Each acre of land would contain the same number of people and the same mix of productive activities. The crucial point in establishing this result is that constant returns

permit each productive activity to be carried on at an arbitrary level without loss of efficiency. Furthermore, all land is equally productive and equilibrium requires that the value of the marginal product, and hence its rent, be the same everywhere. Therefore, in equilibrium, all the inputs and outputs necessary directly and indirectly to meet the demands of consumers can be located in a small area near where consumers live. In that way, each small area can be autarkic and transportation of people and goods can be avoided.

Such an economic space is the quintessence of self-sufficiency. This suggests, therefore, that the constant returns–perfect competition paradigm is unable to cope with the emergence and growth of large economic agglomerations (Krugman 1995, chap. 1).

Increasing returns in production activities are needed if we want to explain economic agglomerations without appealing to the attributes of physical geography. In particular, the trade-off between increasing returns in production and transportation costs is central to the understanding of the geography of economic activities. Although it has been rediscovered many times (including in recent periods), this idea has been at the heart of the work developed by early location theorists. For example, Lösch ([1940] 1954) stated that:

> We shall consider market areas that are not the result of any kind of natural or political inequalities but arise through the interplay of purely economic forces, some working toward concentration, and others toward dispersion. In the first group are the advantages of specialization and of large-scale production; in the second, those of shipping costs and of diversified production (p. 105 of the English translation).

It is only during the 1990s that some trade theorists became aware that "they were doing geography without knowing it" and have turned their attention to spatial issues. Since then, it is fair to say that they have contributed significantly in promoting geographical economics through the use of models involving both monopolistic competition and increasing returns (Krugman 1991a,b; Venables 1996; Helpman 1998).[10]

1.3.2 Agglomeration and Externalities

According to A. Marshall ([1890], 1920, chap. X), externalities are crucial in the formation of economic agglomerations and generate something like a lock-in effect:

> When an industry has thus chosen a location for itself, it is likely to stay there long: so great are the advantages which people following the same skilled trade get from near neighbourhood to one another. The mysteries of the trade become no mysteries; but are as it were in the air, and children learn many of them unconsciously. Good work is rightly appreciated, inventions and improvements in machinery, in processes and the general organization of the business have their merits promptly discussed: if one man starts a new idea, it is taken up by others and combined with suggestions of their own; and thus it becomes the source of further new ideas (p. 225).

For this author, relevant externalities for the formation of clusters involve the following:

1. mass production (the internal economies that are identical to scale economies at the firm's level);
2. availability of specialized input services;
3. formation of a highly specialized labor force and the production of new ideas, both based on the accumulation of human capital and face-to-face communications; and
4. the existence of modern infrastructure.[11]

Despite its vagueness, the concept of Marshallian externalities has been much used in the economics and regional science literature devoted to the location of economic activities because it captures the idea that an agglomeration is the outcome of a "snowball effect" in which a growing number of agents want to congregate to benefit from a larger diversity of activities and a higher specialization.[12] Such cumulative processes are now associated with the interplay of pecuniary externalities in models combining increasing returns and monopolistic competition (Matsuyama 1995).[13]

In fact, the concept of externality has been used to describe a great variety of situations. Following Scitovsky (1954), it is now customary to consider two categories: "technological externalities" (also called spillovers) and "pecuniary externalities." The former deals with the effects of nonmarket interactions that are realized through processes directly affecting the utility of an individual or the production function of a firm. In contrast, pecuniary externalities are by-products of market interactions: They affect firms or consumers and workers only insofar as they are involved in exchanges mediated by the price mechanism. Pecuniary externalities are relevant when markets are imperfectly competitive, for when an agent's decision affects prices, it also affects the well-being of others.

According to Anas et al. (1998), cities would be replete with technological externalities. The same would hold in local production systems (Pyke et al. 1990, chap. 4). In fact, much of the competitiveness of individuals and firms is due to their creativity, and thus economic life is creative in the same way as are the arts and sciences. Of particular interest for creativity are "communication externalities." This idea accords with the view of Lucas (1988, 38) when he writes that "New York City's garment district, financial district, diamond district, advertising district and many more are as much intellectual centers as is Columbia or New York University." Thus, to explain geographical clusters of somewhat limited spatial dimension such as cities and highly specialized industrial and scientific districts, it seems reasonable to appeal to technological externalities, which, in terms of modeling, have the additional advantage of being compatible with the competitive paradigm.

The advantages of proximity for production have their counterpart on the consumption side. For example, the propensity to interact with others is a fundamental human attribute, as is the tendency to derive pleasure in discussing and exchanging ideas with others. Distance is an impediment to such interactions, and thus cities are the ideal institution for the development of social contacts. Along the same line, Akerlof (1997) argued that the inner city is often the substratum for the development of social norms such as conformity and status seeking that govern the behavior of groups of agents.

On the other hand, when we consider a large geographical area, it seems reasonable to think that direct physical contact provides a weak explanation of interregional agglomerations such as the "Manufacturing Belt" in the United States and the "Blue Banana" in Europe (an area that stretches from London to northern Italy and goes through part of western Germany and the Benelux countries). This is the realm of pecuniary externalities that arise from imperfect competition in the presence of market-mediated linkages between firms and consumers and workers. Such externalities lie at the heart of models of monopolistic competition recently developed to explain the agglomeration of economic activities; they also have one major intellectual advantage.

To a large extent, technological externalities are often black boxes that aim at capturing the crucial role of complex nonmarket institutions whose role and importance are strongly stressed by geographers and spatial analysts (see, e.g., Pyke et al. 1990; Saxenian 1994). By contrast, because pecuniary externalities focus on economic interactions mediated by the market, their origin is clearer. In particular, their impact can be traced back to the values of fundamental microeconomic parameters such as the intensity of returns to scale, the strength of firms' market power, the level of barriers to goods, and factor mobility.

Whatever externalities are at work, prices do not fully reflect the social values of goods and services, and thus market outcomes are likely to be inefficient. The dominant feeling in the economics profession is that most cities and agglomerations are just too big. The prevalence of big and gloomy slums in Third World megalopolises gives the impression that the laissez-faire policy has led to an excessive concentration of human beings in excessively large agglomerations all over the world. Likewise, most regional policy debates in industrialized countries implicitly assume that there is too much spatial concentration. In this respect, Hotelling (1929, 57) stated more than 70 years ago what probably remains the conventional wisdom of economists regarding cities and the spatial organization of economic activities: "Our cites become uneconomically large and the business districts within them are too concentrated." We will see in this book that things are not that simple. Urban externalities are not necessarily negative, and increasing returns might be a strong force in favor of geographical concentration. Hence, it seems fair to say that there is no presumption regarding the direction in which governments should move in their regional and urban policies.[14]

1.3.3 Thünen and Agglomerations

At this stage, it is worth noting that the economics profession has ignored the previous availability in Thünen's work of most of the factors explaining economic agglomerations.[15] When asking whether industrial firms are better off located in major cities (especially in the capital), Thünen ([1826] 1966) started by describing the main centrifugal forces at work:

> 1. Raw materials are more expensive than in the country towns on account of the higher cost of transport. 2. Manufactured articles incur the cost of haulage to the provincial towns when they are distributed to the rural consumers. 3. All necessities, especially firewood, are much more expensive in the large town. So is rent for flats and houses, for two reasons (1) construction costs are higher because raw materials have to be brought from a distance and are consequently more expensive, and (2) sites that may be bought for a few thalers in a small town are very dear. Since food, as well as fuel and housing, cost so much more in the large town, the wage expressed in money, must be much higher than in the small one. This adds appreciably to production costs (pp. 286–7 of the English translation).

This list is surprisingly comprehensive. In particular, the impact of high land rents and high food prices on monetary wages in large cities is explicitly spelled out (see Chapter 6).

Thünen then turned to the centripetal forces that, according to him, stand behind industrial agglomerations.

> 1. Only in large-scale industrial plants is it profitable to install labour-saving machinery and equipment, which economise on manual labour and make for cheaper and more efficient production. 2. The scale of an industrial plant depends on the demand for its products. . . . 4. For all these reasons, large scale plants are viable only in the capital in many branches of industry. But the division of labour (and Adam Smith has shown the immense influence this has on the size of the labour product and on economies of production) is closely connected with the scale of an industrial plant. This explains why, quite regardless of economies of machine-production, the labour product per head is far higher in large than in small factories. . . . 7. Since it takes machines to produce machines, and these are themselves the product of many different factories and workshops, machinery is produced efficiently only in a place where factories and workshops are close enough together to help each other work in unison, i.e. in large towns. . . . Economic theory has failed to adequately appreciate this factor. Yet it is this which explains why factories are generally found communally, why, even when in all other respects conditions appear suitable, those set up by themselves, in isolated places, so often come to grief. Technical innovations are continually increasing the complexity of machinery; and the more complicated the machines, the more the factor of association will enter into operation (pp. 287–90 of the English translation).

Observe that the combination of Thünen's agglomeration factors 1, 2, and 4 almost coincides with Krugman's "basic story" for the emergence of a core–periphery structure (see Chapter 9). Furthermore, if we combine these factors

with the last one (7), which is about interindustry linkages and technological spillovers, we get another fundamental explanation for the emergence of industrial agglomerations (see Chapters 7 and 9).

Even though Thünen's work took place at the very beginning of the Industrial Revolution in Germany, it would be hard to imagine a more explicit description of the forces shaping the industrial landscape.

1.4 ON THE RELATIONSHIP BETWEEN SPACE AND ECONOMICS

It is rare to find an economics text in which space is studied as an important subject – if it is even mentioned. As argued by Krugman (1995, chap. 1), this is probably because economists lacked a model embracing both increasing returns and imperfect competition, the two basic ingredients of the formation of the economic landscape, as shown by the pioneering work of Hotelling (1929), Lösch (1940), Isard (1956), Koopmans (1957), and Greenhut (1963).[16]

Certainly many eminent economists have turned their attention to the subject at least in passing, and Samuelson (1983) places the subject's founder, Thünen, in the pantheon of great economists. Thünen ([1826] 1966) sought to explain the pattern of agricultural activities surrounding a typical city in preindustrial Germany, and we will see that his theory has proven to be very useful in studying land use when economic activities are perfectly divisible. In fact, the principles underlying his model are so general that Thünen can be considered the founder of marginalism (Samuelson 1983; Nerlove and Sadka 1991). Ekelund and Hébert (1999, 246) go one step further when they claim that "With uncommon brilliance and deftness Thünen virtually invented the modern economic 'model,' which integrates logical deduction with factual experiment." In addition, the import of Thünen's analysis for the development of geographical economics is twofold in that space is considered as both an economic good and as the substratum for economic activities, thus making his work more relevant and general than several later contributions.

Despite his monumental contribution to economic thought, Thünen's ideas languished for more than a century without attracting widespread attention. Why was this so? According to Ekelund and Hébert (1999, 245), the reason lies in the work and influence of Ricardo:

> The economics of David Ricardo constituted a negative watershed in the history of spatial theory. By reducing situational differences to differences in the fertility of land, Ricardo effectively eliminated spatial considerations from his analytical system. Moreover, he made transportation costs indistinguishable from other costs, and in international trade theory where spatial considerations had previously dominated, he substituted comparative costs as the crucial factor. The practical effect of Ricardo's method and of his analytical innovations was to dislodge space from mainstream economic theory, so that for a long period thereafter it came to be treated, if at all, outside the mainstream deductive models of British classical economics.

Aside from such an unfortunate historical whim, Thünen's theory left a crucial issue unexplored: Why is there a city in Thünen's isolated state? Although such a center may emerge under constant returns when space is heterogeneous (Beckmann and Puu 1985), a city is more likely to arise when increasing returns are at work in the design of trading places or in the production of some goods. In other words, one must appeal to "something" that is not in the Thünian model to understand what is going on.

There is an interesting analogy between the Thünen's model and Solow's (1956) growth model. Both assume constant returns to scale and perfect competition. As in Thünen's, in which the city cannot be explained within the model, the main reason for growth, that is, technological progress, cannot be explained within the model of exogenous growth. This difficulty is well summarized by Romer (1992, 85–6) in the following paragraph:

> The paradox . . . was that the competitive theory that generated the evidence was inconsistent with any explanation of how technological change could arise as the result of the self-interested actions of individual economic actors. By definition, all of national output had to be paid as returns to capital and labor; none remained as possible compensation for technological innovations. . . . The assumption of convexity and perfect competition placed the accumulation of new technologies at the center of the growth process and simultaneously denied the possibility that economic analysis could have anything to say about this process.

Stated differently, explaining city formation in Thünian models is similar to explaining technological progress in the neoclassical growth model.

Despite this limitation, the Thünian model has proven its relevance lately for the development of spatial economics. Following the suggestion made by Isard (1956, chap. 8), Alonso (1964) succeeded in extending Thünen's central concept of bid rent curves to an urban context in which a marketplace is replaced by an employment center (the Central Business District). Since that time urban economics has advanced rapidly. The reason for this success is that the model is compatible with the competitive paradigm. Or, as pointed out by Krugman (1995, 54),

> Economists understood why economic activity spreads out, not why it becomes concentrated – and thus the central model of spatial economics became one that deals only with the way competition for land drives economic activities away from a central market.

1.4.1 Space and the Competitive Paradigm

More than half a century ago, when Isard (1949) critically discussed general equilibrium analysis, he was mainly concerned with Hicks's *Value and Capital* published in 1939. Isard concluded that Hicks confined himself to "a wonderland of no dimension." He further elaborated this point on page 477 in which he recorded a conversation he had with Schumpeter, who defended the Hicksian analysis, maintaining that "transport cost is implicitly contained in production

cost, and thus Hicksian analysis is sufficiently comprehensive." Isard's point was that

> production theory ... cannot justifiably treat certain production costs explicitly and other important ones implicitly in order to avoid the obstacles to analysis which the latter present. For a balanced treatment, the particular effects of transport and spatial costs in separating producers from each other must be considered.[17] They are too vital to be sidestepped through implicit treatment, as Hicks and others may be interpreted as having done.

We believe that Isard was right.

In fact, the debate about whether or not the general equilibrium model based on perfect competition is comprehensive enough to fully reflect the working of the spatial economy has a long history. On one side, general equilibrium theorists have maintained that the problem of space can be handled by defining each commodity by its physical characteristics as well as by the place (period) in which it is made available, and hence, once we have thus indexed commodities, we can essentially forget space (and time) in economic theory. This is the way Arrow and Debreu (1954) treated space (and time) in their seminal article.

On the other side, from the standpoint of the alternative view, supported by Lösch, Isard, and several others, the problem is not that simple. To capture the essential impact of space on the distribution of economic activities, new models are needed that are fundamentally different from those found in standard general equilibrium. In particular, Koopmans claimed in his *Three Essays on the State of Economic Science* that the vital effects of space become evident when our concern is the location of several economic activities and, hence, when the spatial distribution of activities itself becomes a variable. In this respect, Koopmans (1957, 154) maintained that

> without recognizing indivisibilities – in human person, in residences, plants, equipment, and in transportation – urban location problems, down to those of the smallest village, cannot be understood.

Because standard general equilibrium analysis abstains from the consideration of indivisibilities or increasing returns to scale, it will fail to capture the essential impact of transport and land when one comes to study the spatial distribution of economic activities.

In the long debate concerning the comprehensiveness of general equilibrium theory for the spatial economy, Starrett (1978) has made a fundamental contribution. The essential question is whether the competitive price mechanism is able to explain the endogenous formation of economic agglomerations. To check the ability of a spatial model to do so, the best approach is to consider the case of a homogeneous space in which economic agents are free to choose their locations. For, if any concentration of economic activities is to occur, it must be due to endogenous economic forces. Starrett has shown that if space is

homogeneous and transport costly, then any competitive equilibrium is such that no transportation occurs. In other words, the economy degenerates into separated single-location groups of agents with all trades taking place within, rather than between, groups. Consequently, the perfectly competitive price mechanism alone is unable to deal simultaneously with cities and trade. This fact has a fundamental implication for the modeling of the spatial economy: If the purpose is to build a theory explaining the formation of economic agglomerations, then such a theory must depart from general competitive analysis.

Once it is recognized that the competitive equilibrium paradigm cannot be the right foundation for the space-economy, what theory is conceivable? The following is Isard's second major insight to which the alternative should be a general theory of spatial competition:

> Because of the monopoly elements which are almost invariably present in spatial relations, a broadly defined general theory of monopolistic competition can be conceived as identical with the general theory of location and space-economy (Isard 1949, 504–5).[18]

1.4.2 Spatial Competition

Ever since Sraffa (1926), it has long been recognized in the economics profession that the integration of increasing returns within the competitive model is problematic. As observed by B.C. Eaton and Lipsey (1977, 63),

> Once the firm acts as if it faces a perfectly elastic demand curve, there is nothing to restrict size from the demand side. *Size must be restricted from the cost side*. Hence, the extreme importance of eventually diminishing returns to scale in any competitive model that seeks to limit the size of plants and firms. (italics in original)

while, despite unexploited economies of scale, however, the firm size is demand constrained once consumers are dispersed across locations.

In fact, combining space and economies of scale has a profound implication for economic theory. If production involves increasing returns, a finite economy accommodates only a finite number of firms, which are *imperfect competitors*. Treading in Hotelling's footsteps, Kaldor (1935) argued that space gives this competition a particular form. Because consumers buy from the firm with the lowest price augmented by transport cost, each firm competes directly with only a few neighboring firms regardless of the total number of firms in the industry (Eaton and Lipsey 1977; Gabszewicz and Thisse 1986).

The very nature of spatial competition is, therefore, oligopolistic and should be studied within a framework of interactive decision making. This was one of the central messages conveyed by Hotelling (1929) but was ignored until economists became fully aware of the power of game theory for studying competition in modern market economies.[19] Following the application of game theory to industrial organization in the late 1970s, it became natural to study the implications of space for competition. New tools and concepts are now available to revisit and formalize the questions raised by early location theorists.

But this is not yet the end of the story. Most of the contributions to location theory by industrial organization deal with partial equilibrium models. Although a comprehensive general equilibrium model with imperfect competition has so far been out of reach and is likely to remain so for a long time (Bonanno 1990), specific models have been developed that, taken together, have significantly improved our understanding of how the spatial economy works. In particular, since the 1990s, a growing number of economists have become interested in the study of location problems, and it is fair to say that some real progress has been made. This increased interest has been partially triggered by the integration of national economies within trading blocks, such as the European Union or NAFTA, that leads to the fading of national borders. In the same vein, the study of the microeconomic underpinnings of economic development has led several economists to investigate the connection between growth and cities.[20]

1.5 PLAN OF THE BOOK

To a large extent, the organization of this book reflects what we have said in the foregoing sections. Although we have tried to make each chapter more or less self-contained, the reader may benefit from "agglomeration economies" in the course of study. Thus, the book has been organized into four parts. The first one deals with the fundamentals of geographical economics. After showing the insufficiency of the competitive paradigm for studying economic geography, we consider different issues such as the land rent formation, the structure of competition between geographically separate firms, and the provision and financing of local public goods. The second part explains the structure of metropolitan areas and the clustering of firms selling similar products. In the third part, we shift to a different geographical scale and cope with the impact of factor mobility on the location of industry. In particular, we study the role of both technological and pecuniary externalities in the interregional distribution of firms. In the last part, we offer two syntheses of various approaches taken in this book, which also suggest new lines of research. We first study how perfect competition in the land market and monopolistic competition in the product market can be combined with the aim of explaining the emergence of cities in an otherwise homogeneous setting. We then proceed by investigating the relationship between agglomeration and growth once agents have forward-looking behavior.

Needless to say, the topics covered in this book reflect our idiosyncrasies. Hence, we owe our apologies to those who have contributed to the field but who might dislike our choice of menu.

1.5.1 Part I. Fundamentals of Geographical Economics

Chapter 2 shows the insufficiency of the competitive paradigm for the formation of economic agglomerations. Specifically, we follow Starrett and show

that cities, local specialization, and trade cannot arise at the competitive equilibrium of an economy with a featureless space. This criticism, because it is internal to the model, is especially powerful. After having provided an intuitive explanation for Starrett's theorem, we discuss what could be the alternative modeling strategies that will allow us to study economic agglomerations in market economies.

Chapter 3 discusses the location of divisible activities, as directly inspired by Thünen, and demonstrates how a competitive land market works regarding the allocation of land among competing activities. Once it is assumed that centers do exist through which commodities are traded, the competitive model is applicable and yields sensible results regarding the way land is organized around these centers. (The reasons for the formation of centers are postponed to Part II.) We then consider the adaptation of the Thünian model to urban economics. The main results derived in the classical context of the monocentric city are then presented.

In Chapter 4, we move to the fundamental trade-off between increasing returns and transport costs and investigate several models illustrating the importance of this trade-off for the spatial economy. Our first task is to explain why the gathering of people within a small area is able to yield scale economies in the aggregate. We consider two microeconomic foundations for such social returns. In the first, a monopolistic, competitive, intermediate sector produces nontradable goods under scale economies at the firms' level. Increasing returns are transferred in the aggregate to the final sector that would otherwise exhibit constant returns, thus showing the importance of the urban service basis for the formation and productivity of the city production system. In the second model, both firms and workers are heterogeneous, whereas wage formation is driven by a matching process. The average quality of the match rises with the population size, and this factor suggests an explanation for the tendency of wages to be higher in large metropolitan areas than in smaller cities.

We then focus on the process of competition among communities (e.g., company towns) that form to exploit scale economies. When communities are able to capitalize land rent into their payoffs, we show that increasing returns do not prevent the decentralization of the optimal allocation. Production communities, on the contrary, do not form in the absence of increasing returns. This material allows us to present the basic elements of a theory of urban systems proposed by Henderson.

Finally, we demonstrate how the process of spatial competition develops once it is recognized that geographical separation gives firms market power over consumers located in their vicinity. If firms are able to capitalize the land rent they create by their mere existence, then they find it profitable to sell at marginal costs, making money from the land rent only. An old conjecture stated by Hotelling (1938) is then proven: When there is free entry, firms' fixed costs are just covered by the aggregate land rent.

A similar line of reasoning is used in Chapter 5 but is applied to local public goods. Our first result is the Henry George theorem, which claims that public expenditure equals aggregate land rent when the population size of a city is optimal. In the same spirit as in Chapter 4, we show how competition among land developers allows for the decentralization of the efficient allocation of public goods when agents are identical in preferences and incomes. We also consider voting as an alternative decision-making mechanism to determine the location and number of facilities supplying local public goods. It is shown that voting fosters too much public infrastructure financed through too big a public budget. Once again, allowing individuals to move and to compete for land use permits us to show that the optimum is unanimously selected by consumers through voting. All these results confirm the idea that a competitive land market is a powerful device for improving the allocation of resources.

1.5.2 Part II. The Structure of Metropolitan Areas

In Chapters 6 and 7, we deal explicitly with the formation of different types of economic agglomerations within cities. Specifically, we survey and extend the literature developed in urban economics, industrial organization, and regional science to explain either the emergence of a central business district or the clustering of firms selling similar products. In Chapter 6, our frame of reference is the existence of communication externalities. We first consider partial equilibrium models, the aim of which is to determine under which conditions similar agents (households or firms) want to congregate despite their competition for land. We show that the density around the endogenous center is not high enough from the welfare point of view because each agent accounts for the benefit received from the others but not for the benefits transmitted to others.

We then move to an explicit treatment of spatial interaction between firms and households on both land and labor markets. The urban structure turns out to be the outcome of the interplay between the intensity of face-to-face communications and the level of commuting costs. Low commuting costs foster the emergence of a single central business district, thus providing a key explanation for the monocentric city. However, dispersed or polycentric structures may emerge when higher commuting costs prevail. Typically, a multiplicity of equilibria arise, and transitions from one equilibrium to another may display catastrophic changes.

Chapter 7 focuses on imperfect competition as the main explanation for the clustering of firms within cities. Without a strictly positive markup generated by product differentiation, there would be no agglomeration in the models analyzed. We deal with the case of monopolistic and oligopolistic competition and consider mobile as well as fixed consumers. As observed by Stahl (1983), product variety is a major determinant of consumers' spatial behavior. The

general message of this chapter is that low transport costs together with sufficient product differentiation push economic agents toward agglomeration. The reason is that product differentiation relaxes price competition and consequently allows firms to attract more consumers when they are clustered than when any firm chooses to stand alone.

1.5.3 Part III. Factor Mobility and Industry Location

In Chapters 8 and 9, our interest shifts from cities to the spatial distribution of industries among larger spatial entities, that is, regions or nations. Hence, land consumption is no longer an issue. This is not to deny the reality of congestion effects, but we believe that they have little to do with the imbalance between big regions. At this geographical scale, the reasons for over- or underconcentration have more to do with interindustry linkages or linkages between firms and consumers, workers, or both, through the product and labor markets. Chapters 8 and 9 can be viewed as the counterpart of Chapters 6 and 7, respectively, because externalities and imperfect competition are the corresponding engines of agglomeration in each pair of chapters. Their aim is to present "clarifying examples" enhancing our understanding of how the obstacles to the spatial mobility of goods and factors affect the economic geography. In particular, these chapters provide illustrations of what is likely the main spatial feature of modern economies, namely, the emergence of a "putty-clay" economic geography. Specifically, the recent fall in trade costs seems to allow for a great deal of flexibility in where particular activities can locate, but once spatial differences have developed, they tend to become rigid. Hence, regions that were once similar may end up having very different production structures.

Chapter 8 is devoted to the impact of technological externalities, whereas Chapter 9 is concerned with pecuniary externalities expressed through monopolistic competition. In Chapter 8, we deal with the existence of urbanization economies in an otherwise standard model of regional economics. The sole presence of such externalities suffices to upset the convergence result derived in the standard neoclassical model. We then shift to localization economies to investigate the interplay between the fall in trade costs, the cost reductions associated with the implicit cooperation arising among firms located in the same area, and the intensity of competition between firms in the domestic and foreign product markets. When trade costs keep falling, an asymmetric distribution of firms emerges gradually from the interplay between these three forces.

Chapter 9 deals with what has come to be known as the "new economic geography." What drives the formation of agglomeration here is the presence of many types of pecuniary externalities such as those created by firms or workers moving from one region to the other. We will restrict ourselves to the description of the main forces driving the core–periphery structure, namely, when preference for variety and increasing returns combine to generate

economic agglomerations.[21] In doing so, we compare two alternative formulations of preferences (CES versus quadratic utilities) and of transport technologies (iceberg versus proportional costs). We also study the impact of the intermediate sector (as modeled in Chapter 4) on the spatial distribution of firms and provide a welfare analysis of the core–periphery model. Finally, we complete this chapter by extending the standard framework to deal with the issue of "history versus expectation" (Krugman).

1.5.4 Part IV. Urban Systems and Regional Growth

In the final two chapters, we show how the material developed in previous chapters may be used to address two major economic issues: the formation of urban systems and the unequal growth of regions. In Chapter 10, we combine different models studied in previous chapters in order to develop a synthetic approach whose aim is to explain how and why cities emerge as a response to population growth. For that, we graft a competitive land market associated with the agricultural sector onto the canonical core–periphery model studied in Chapter 9. A monocentric configuration arises as a spatial equilibrium when the transport cost of the agricultural good is low relative to the cost of moving the industrial goods and when the total population is small. Using the intermediate input framework developed in the previous chapter leads to a wider array of results. In particular, we show that two very distinct types of monocentric patterns may emerge according to the level of intermediate inputs' transport costs. Finally, we discuss how these models can be used to explain the regular pattern of cities suggested by central place theory when population grows continuously. In addition to their theoretical interests, the results presented in this chapter shed light on the urbanization phases that took place in the United States during the second half of the nineteenth century.

As will be seen in the course of this book, geographical economics has strong connections with several branches of modern economics, including industrial organization and urban economics but also with the new theories of growth and development. In particular, economic geography and endogenous growth theory share the same framework, using monopolistic competition, increasing returns, and spillovers. This suggests the existence of a high potential for cross-fertilization. Indeed, regional growth turns out to be a new and promising topic, although it is still in its infancy.

In Chapter 11, we deal with some of the main issues addressed in the hope of convincing the reader of the relevance of further research in this domain. The main message here seems to be that, in a world of globalization, agglomeration may well be the territorial counterpart of economic growth much in the same way as growth seems to foster inequality among individuals. However, inequalities may be accompanied by a higher level of welfare even for those living on the periphery. If such preliminary results were to be confirmed, they would have

farfetched implications for the modern space-economy as well as for the design of more effective economic policies.

NOTES

1. The term *agglomeration* is less ambiguous than *concentration*, which is used to describe different economic phenomena. *Agglomeration* has been introduced in location theory by Weber ([1909] 1929). Though Weber is mainly known for his work on the location of the firm (Wesolowsky 1993), his main concern was to explain the formation of industrial clusters (Isard 1956, chap. 2).
2. In this respect, R. Martin (1999, 387) is right in his criticism of economists' proclivity to use the same models "to explain the tendency for economic activity to agglomerate at various spatial scales, from the international, through the regional, to the urban and the local."
3. Before proceeding, we would like to clarify how this book relates to two recent volumes. First, the present book differs essentially from the work by Fujita, Krugman, and Venables (1999), which focuses exclusively on monopolistic competition à la Dixit–Stiglitz. In contrast, we consider a broader range of approaches and concepts, with a special focus on cities, in order to study the foundations of the spatial economy. We also cover more broadly the economics and regional science literature that have been devoted to the location of economic activities. The two books are therefore complementary, defining the frontier of geographical economics. Our book also differs from the one edited by Huriot and Thisse (2000), which deals more with various specific urban issues (e.g., the dynamics of cities when land is not malleable) or particular aspects of the process of agglomeration (e.g., the impact of globalization on the geography of financial centers) that are not covered here. Once again, there is complementarity.
4. Throughout this book, the word *city* refers to a whole urban region; we will use *city*, *metropolitan area*, and *urban area* interchangeably.
5. Throughout this book, transportation costs are broadly defined to include all impediments caused by distance such as shipping costs per se, tariff and nontariff barriers to trade, different product standards, difficulty of communication, and cultural differences.
6. It is worth noting here that Isard and Peck (1954) tried to echo Ohlin's concern about the relevance of transport costs in trade theory. Isard (1954) also strove to provide an early justification for the gravity model, which was familiar to Tinbergen as well (1962).
7. This is what Cronon (1991) calls "first nature" by contrast to "second nature," which emerges as the outcome of human beings' actions to improve upon the first one.
8. See Razin and Sadka (1997) for a synthetic presentation of migration and trade as possible substitutes.
9. It has recently been argued that capital does not necessarily flow from rich to poor regions (Lucas 1990), whereas persistent regional wage differences seem to be frequent within modern economies (Shields and Shields 1989). In addition, the empirical evidence that per capita income would converge across countries, or even

between regions of the same country, is not conclusive. Without being complete, we should like to mention Sala-i-Martin (1996), Blanchard and Katz (1992), de la Fuente and Vives (1995), de la Fuente (1997), and Quah (1996).

10. In the 1970s, another prominent trade theorist, R.G. Lipsey, vastly contributed, with B.C. Eaton, to the development of spatial economic theory (see, e.g., Eaton and Lipsey 1977, 1997).
11. An attempt to clarify the concept of Marshallian externalities is made in Section 4.2.
12. This phenomenon is similar to that encountered in network externalities. Besides the network effect, which is an agglomeration force because consumers always prefer a larger network, it is necessary to identify another effect that plays the role of a dispersion force in order to obtain different networks (see Grilo, Shy, and Thisse 2001 for a spatial model with network externalities). Note also that the issue of standardization bears some resemblance to that of agglomeration (Arthur 1994, chaps. 2 and 4).
13. In a sense, this corresponds to a revival of ideas advocated by early development theorists who used various related concepts such as the "big push" of Rosenstein-Rodan (1943), the "growth poles" of Perroux (1955), the "circular and cumulative causation" by Myrdal (1957), and the "backward and forward linkages" by Hirschman (1958). Recent additions to this cornucopia include the "dynamic economies of scale" by Kaldor (1985), the "positive feedbacks" by Arthur (1994, chap. 1) and the "complementarities" by Matsuyama (1995).
14. The idea that cities have an optimal size is old and goes back at least to Plato, for whom the ideal city has 5,040 citizens. This number does not include women, children, slaves, and foreigners, thus making the total number of residents significantly larger (we thank Yorgos Papageorgiou for having pointed out this reference to us).
15. See section 2 of part II of *The Isolated State*, which contains the extracts of posthumous papers on location theory written by Thünen between 1826 and 1842 and edited by Hermann Schumacher in 1863. The reader is referred to Fujita (2000) for more details.
16. See Ponsard (1983) for a historical survey of spatial economic theory.
17. It is not clear what Isard meant here by "the particular effects of transport and spatial costs in separating producers from each other." But, because Isard complained in the same paper about Hicks's rejection of monopolistic competition model in favor of perfect competition, we guess that "the particular effects" include the monopolistic elements that spatial costs introduce into price theory.
18. Of course, Isard does not refer here to the Dixit–Stiglitz model of monopolistic competition but more broadly to what is now called imperfect competition.
19. In this article, Hotelling's contribution to economic theory has been fundamental in many respects. For example, Mueller (1989, 180) regards Hotelling's paper as the pioneering contribution in public choice. The idea to formulate a game on price and locations according to a two-stage procedure was also extremely ingenious and original; it precedes by several decades the work of Selten on perfect equilibrium.
20. It is worth noting that preclassical economists have stressed the role of cities in the process of development and growth (see, e.g., Lepetit 1988, chap. 3, for an overview of the main contributions before Adam Smith). In particular, those economists viewed cities not only as a combination of inputs but also as a "multiplier" that leads

to increasing returns in the aggregate. In accord with modern urban economics, preclassical economists further considered cities as economic agents having the power to make decisions. Not surprisingly, their work is connected to modern theories of growth, thus suggesting that the "new" theories of agglomeration and of endogenous growth have the same historical roots. There are here several interesting questions that should be explored by historians of economic thought.

21. Using product variety as a surrogate for urban life agrees with the early work by Cantillon (1755). According to this author, the origin of cities was to be found in the concentration of land ownership, allowing landowners to live at a distance from their estates in places where they could "enjoy agreeable society," and in an agglomeration economy related to the landowners' demand, which attracted craftsmen and merchants.