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