# 1

# Agglomeration and Economic Theory

## 1.1 INTRODUCTION

This book is an attempt to uncover the main economic reasons for the existence of peaks and troughs in the spatial distributions of population and wealth. Economic activities are not concentrated on the head of a pin, nor are they spread evenly over a featureless plain. On the contrary, they are distributed very unequally across locations, regions, and countries, generating contour lines that vary with time and place. 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. Furthermore, paralleling large and small planets, there are large and small settlements with very different combinations of firms and households. Though universal, these phenomena are still in search of a general theory.

The common belief, however, is that we now live in a world where the tyranny of distance, which has been such a powerful force in human history, no longer exists. The spectacular and steady drop in transport costs since the mid-nineteenth century, compounded by the decline of protectionism and, more recently, by the near disappearance of communication costs, is said to have freed economic agents from the need to be located near one another, suggesting 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 vanish. In other words, the combined impact of technology and globalization would make the traditional geography of economic activity obsolete and yesterday's world of peaks and troughs would miraculously become "flat."

Recent empirical and theoretical work in new economic geography and urban economics shows a very different reality. While it is true that the importance of proximity to natural resources has declined considerably, thus giving firms and households more freedom to locate where they wish, this does not mean that distance and location have disappeared from economic life. Quite the contrary, economic geography points to new forces, hitherto outweighed by

natural factors, which are shaping an economic landscape that, with its many barriers and large inequalities, is anything but a "flat world." For example, the simplistic view that improvements in communication technologies will render financial centers obsolete is as misplaced as the opposite view that we will see ever-increasing concentration. As is seen throughout this book, the huge drop in transport and communication costs is precisely what allowed these new forces to emerge and subsequently fostered the greater productivity that now characterizes large cities.

Ever since the emergence of civilization, human activities and standards of living have been unevenly distributed among both the continents and their territories. According to the famous historian Braudel (1979, 39), a "world-economy" is the combination of three types of space:

> The centre or core contains everything that is most advanced and diversified. The next zone possesses only some of these benefits, although it has some share in them: it is the "runner-up" zone. The huge periphery, with its scattered population, represents on the contrary backwardness, archaism, and exploitation by others.

Even though economic activities are, to some extent, spatially concentrated because of natural features (think of rivers and harbors), it is our contention that economic mechanisms yielding agglomeration (or dispersion) of activities are more fundamental. As is discussed in this book, most of these mechanisms rely on the fundamental *trade-off* between various forms of *increasing returns* and different types of *mobility costs*. It applies to different types of scale economies and to the costs generated by the transfer of people, goods, and information. In addition, this trade-off is valid on all spatial scales (cities, regions, countries, or continents), which makes it a valuable analytic tool.

Although referring to agglomeration as a generic term is convenient, 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 the North-South divide. High-income nations are still clustered in small industrial cores in the Northern Hemisphere whereas productivity per capita steadily declines with distance from these cores.

Economic growth has always been and still is geographically unequal and very localized. This is especially well illustrated by the emergence of a core-periphery structure in Europe during the nineteenth century (Bairoch 1997). From 1800 to 1913, European countries experienced a high rate of growth. Yet, while the initial levels of economic development were roughly the same, varying by about 10 percent around the European average, countries were affected quite differently by the Industrial Revolution and its concomitant decrease in transport costs. Indeed, international differences grew progressively and

[1] The term "agglomeration" is less ambiguous than "concentration," which is used to describe different economic phenomena.

reached a ratio of 1 to 4 between the richest and poorest nations by 1913. While the average European GDP per capita increased gradually by a factor slightly greater than 2.5, the standard deviation increased even faster, going from 24 in 1800 to 229 in 1913. In other words, disparities between nations grew more than proportionally, the coefficient of variation increasing from 0.12 in 1800 to 0.42 in 1913. As usual, such aggregate measures hide even stronger contrasts between countries: while the GDP per capita of the United Kingdom increased by a factor exceeding 4, that of the Balkans (Bulgaria, Greece, and Serbia) barely rose by 50 percent.

A more recent example of such a phenomenon is provided by the evolution of world production during the last two decades of the twentieth century. We focus on the three biggest regional blocks, that is, the European Union (EU), the North American Free Trade Agreement (NAFTA), and East Asia.[2] At the global scale, in 1980 the EU-15 accounted for 29 percent of world gross domestic product (GDP), NAFTA for 27 percent, and East Asia for 14 percent. These three blocks thus produced 70 percent of world GDP. Twenty years later, the share of the EU-15 had fallen to 25 percent, while that of NAFTA had increased to 35 percent and that of East Asia to 23 percent. Together, they accounted for 83 percent of world GDP in 2000, a much larger share than in 1980. Throughout this period, we saw more agglomeration and rapid growth of the world economy with significant reductions in transport and communication costs, which fostered the international division of labor. In particular, East Asia has emerged as the world's manufacturing center (e.g., it provided 100 percent of digital cameras and hard disk drives, 99.8 percent of personal computers, and 71.5 percent of cell phones in 2008).

At the national level, a few large cities produce a sizable share of their country's GDP. For example, in the Republic of Korea, the capital region, which covers 11.8 percent of the country's surface area and includes 48.6 percent of the population, produced 47.8 percent of Korea's GDP in 2008. In France, the metropolitan area of Paris, which accounts for 2.2 percent of the country's area and 18.2 percent of its population, produced 28.3 percent of its GDP. Similarly, in Brazil, the world's fifth-largest country in surface area, 33.9 percent of GDP is produced by 21.6 percent of the population in the state of São Paulo, which occupies only 2.9 percent of the country's area. In 2000, the thirty-eight largest cities of the EU-15 accounted for 27 percent of its jobs and produced 29 percent of its GDP.

Highly diverse size and activity arrangements also exist at the regional and urban levels. Cities may be specialized in a very small number of industries, as are many medium-size cities. However, large metropolises such as New York or Tokyo are highly diversified in that they nest several industries that are not

[2] East Asia here is formed by China, Indonesia, Japan, Hong Kong, the Republic of Korea, Malaysia, the Philippines, Singapore, Thailand, and Taiwan.

related through direct linkages. 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), manifest various types of local specialization.

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 regional or urban policies. What distinguishes these various types of agglomeration is the *spatial scale*, or the spatial unit of reference, chosen in conducting one's research, just as there are different levels of aggregation of economic agents.

The economic reasons that stand behind strong geographical concentrations of consumption and production are precisely what we aim to investigate in this book. To achieve this objective, we appeal to the concepts and tools of modern microeconomics. Even though clusters appear at different geographical scales, it would be futile to look for *the* model explaining different types of economic agglomerations, the reason being these clusters involve distinct and contrasting details. This should not come as a surprise, for geographers have long known that spatial scale matters: What is true at a certain spatial scale is not necessarily true at another – the "ecological fallacy." In this respect, 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."

To study issues arising at, say, the interregional and intra-urban levels, different models encapsulating specific spatial features are developed in this book, the reason being that *the nature and balance of the system of forces at work at different spatial scales need not be the same*. 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."

However, as is seen, working with different models will not prevent us from deriving a few general principles that seem to govern the formation of distinct agglomerations. The main thrust of the book is that a few basic ideas and concepts lie at the foundations of the still needed general theory of location.

In particular, a major principle holds true regardless of the scale of analysis retained: The emergence of economic agglomerations is associated with the emergence of spatial inequalities. Such inequalities are often at the origin of strong tensions between different political bodies or jurisdictions, or even social, religious, or ethnic groups when they are geographically concentrated.

Understanding how spatial inequalities in economic activities and living standards arise is thus a fundamental challenge for economists, regional scientists, and geographers alike.

### 1.2 DO TRANSPORT COSTS STILL MATTER?

By its very nature, transport is linked to trade. Trade being one of the oldest human activities, the transport of commodities is, therefore, a fundamental ingredient of any society. People get involved in trade because they want to consume goods that are not produced within reach. The Silk Road provides evidence that shipping high-valued goods over long distances has been undertaken because of this precise reason.

The transportation sector underwent the most stunning changes during the Industrial Revolution. According to Bairoch (1997, volume 2, 26), "On the whole, between 1800 and 1910, it can be estimated that the lowering of the real (weighted) average prices of transportation was on the order of 10 to 1." For example, prior to the Industrial Revolution the average cost of ground transportation of grains per ton–kilometer was equal to the average cost of buying 4 or 5 kg of grain, but this cost fell to 0.1 kg per ton–kilometer in 1910 because of long-distance transportation by rail. Once we account for the decrease in the price of grain generated by technological innovations in agriculture, the decrease in transport costs is even larger: they are divided by a factor close to 50 (Bairoch 1997, chap. 4). It is, therefore, legitimate to ask whether it is still relevant to pay attention to transport costs in economic models.

Our answer involves two pieces of argument that are very different in nature. First, even though transport costs must be positive for space to matter, one should not infer from this observation that location matters less when transport costs decrease. Quite the opposite, by making them more footloose, new economic geography and urban economics show that lower transport costs make firms more sensitive to minor differences between locations. As a result, *a tiny difference across places may have a big impact on the spatial distribution of economic activity*. Such effects cannot be understood under the assumption of zero transport costs.

Second, transport (or trade or shipment) costs must be defined broadly enough to include all impediments to trade and movements. Spulber (2007) refers to them as "the four Ts": (i) Transport costs per se because goods have to reach their consumption place, whereas many services remain nontradable and various exchanges still require face-to-face contacts; (ii) Time costs as, despite Internet and video conferences, there are still communication impediments across dispersed distribution and manufacturing facilities that slow down reactions to changes in market conditions, whereas the time needed to ship certain types of goods has a high value; (iii) Transaction costs that result from doing business at a distance because of differences in customs, business practices, and

political and legal climates; and (iv) Tariff and non-tariff costs such as different anti-pollution standards, anti-dumping practices, and the massive regulations that still restrict trade and foreign investments.

Furthermore, distribution costs, because of wholesalers and retailers, may be added to transport costs. Anderson and van Wincoop (2004) estimate that, for developed countries, average trade costs represent 170 percent of the produce price of manufactured goods. Trade costs consist of 55 percent internal costs and 74 percent international costs ($2.7 = 1.55 \times 1.74$). The international costs are in turn broken down into 21 percent transport costs and 44 percent costs connected with border effects ($1.74 = 1.21 \times 1.44$). Such results clash with Glaeser and Kohlhase (2004), who observe that the average cost of moving a ton a mile in 1890 was 18.5 cents, as opposed to 2.3 cents today (in 2001 dollars). It should be stressed, however, that Glaeser and Kohlhase focus on the sole transport sector within the United States, whereas Anderson and van Wincoop adopt a much broader perspective. Furthermore, the variance in trade costs across goods is large. Thus, although shipping certain goods is almost costless, trading some others remains very expensive.

Evidently, the relative importance of the four Ts varies enormously from one sector to another, from one activity to another, and from one commodity to another. Understanding the relative evolution of these components is, therefore, needed to figure out how transport costs evolve. Regardless of the elements accounted for, the intensification of competition, which has been caused in part by the secular decrease of transport costs, has made logistics an increasingly important issue in a world where firms are in search of flexibility. Being the inherent attribute of exchanges across places, transportation costs remain central to the formation of trade flows and the location of economic activity.

At a very different spatial scale, commuting is far more costly than what is commonly believed. Twenty kilometers a day to work costs thousands of euros a year (Glaeser 2011). That said, although we acknowledge the importance of understanding the evolution of the elements that determine the level of various types of transport costs, we are content with treating these costs and the transport industry as black boxes.[3]

### 1.3 CITIES: PAST AND FUTURE

Urbanization is probably the most extreme form of geographical unevenness. Indeed, casual observation reveals the extreme variation in the intensity of

[3] There is a need for a stronger connection between, on the one hand, transport economics and, on the other, new economic geography and urban economics. The state of the art in transport economics is provided by de Palma et al. (2011).

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. 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 and economic 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 towns and cities. Historical evidence shows that the existence of cities has increased the efficiency of trade, industry, and government, raising it to a level unattainable with a scattered population. Adam Smith's example of farmers in the Scottish Highlands, who had to work at a large number of different activities to survive, provides a contrary illustration of the validity of this assertion. Once the impact of increasing returns is recognized, cities may be viewed as "economic multipliers" magnifying individual decisions.

Although the sources are dispersed, not always trustworthy, and hardly comparable, data clearly converge to show the existence of an urban revolution, which started with the Industrial Revolution. In Europe, the proportion of the population living in cities increased very slowly from 10 percent in 1300 to 12 percent in 1800 (Bairoch 1988). It was approximately 20 percent in 1850, 38 percent in 1900, 52 percent in 1950, and is now close to 75 percent, thus showing an explosive growth in the urban population. In the United States, the rate of urbanization increased from 5 percent in 1800 to more than 60 percent in 1950 and is now nearly 77 percent. In Japan, the rate of urbanization was about 15 percent in 1800, 50 percent in 1950, and is now about 78 percent. The proportion of the urban population in the world increased from 30 percent in 1950 to 45 percent in 1995. According to the United Nations Population Fund, "In 2008, the world reaches an invisible but momentous milestone: For the first time in history, more than half its human population, 3.3 billion people, will be living in urban areas." 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. In 2012, 26 cities exceed 10 million in population, the top five being in Asia.

[4] Throughout this book, the word *city* refers to a whole urban region; we use city, metropolitan area, and urban area interchangeably.

The largest one, Tokyo, with 37 million, exceeds the second one, Jakarta, by 11 million.

The first towns came into existence in different parts of the world with the creation of an agricultural surplus. Once this had been achieved, the effects of the social division of labor began to be felt and became finer when the geographical concentration of some people arose. Scale economies are always found in one form or another behind the process of urbanization, whether in business activities or in the supply of public goods and services (courts, hospitals, or universities). The pre-industrial town was dominated by the big landowners living there, as well as by the activity of the merchants and artisans. For example, according to Cantillon (1755), the origin of cities is to be found in the concentration of land ownership, which allows landowners to live at a distance from their estates in places where they "enjoy agreeable society," and in the landowners' demand, which attracts craftsmen, who produce nontradable consumption goods and services, and merchants, who buy and carry luxury goods produced in distant places. However, the pre-industrial town was also, and perhaps above all, a marketplace as well as a provider of services for the surrounding countryside.

Towns multiplied and their development undertook a profound change with the Industrial Revolution. Scale economies at the plant level played a fundamental role in this renewal of the urbanization process. The industrial town is typically the location where the means of production are combined under the same roof to take advantage of the division of labor. Initially, these new towns accommodated both workers and the owners of production means. However, the owners of capital increasingly left them for the conveniences of the big cities (as did the landowners before them), which were also both financial and political centers. Beyond the city fringe spreads what Lewis Mumford called the "invisible city," which symbolizes the influence that the visible city exerts over consumption, culture, and life styles, even in the smallest villages, because of dramatic advances in transportation and telecommunications.

Today, the big city is the result of a richer causality, which includes the specialization and diversification of tasks, the widening of the range of choices of consumption goods and production factors, as well as the various communication externalities. The post-industrial city (which may have played a major role in earlier periods; think of London, Milan, and Paris) is associated with the growth in services and, more recently, with its role as a node in communication networks. It is more diversified than the industrial town and thus better able to resist sectoral shocks. Finally, it is the source of most technological and social innovations. While scale economies within firms still play a role, increasing returns in the post-industrial city are perhaps to be found elsewhere, in the form of pecuniary and technological externalities.

One of the general principles 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*, although further decreases may yield dispersion of some activities owing to factor price differentials. In addition, technological progress keeps bringing new types of innovative activities that benefit most from being agglomerated and, therefore, tend to arise in developed and rich areas. Consequently, the wealth or poverty of nations and regions seems to be more and more related to the development of prosperous and competitive clusters of industries as well as to the existence of large and diversified metropolitan areas. 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, especially in the Age of the Internet. Economists, regional scientists, and geographers must explain why firms and households concentrate in large metropolitan areas where housing and commuting costs are higher than in smaller urban areas.

In this book, we intend to address the main causes for the formation of the various types of economic agglomerations previously described. As discussed in the next 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 leads us 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.

## 1.4 WHY DO WE OBSERVE ECONOMIC AGGLOMERATIONS?

The main principles that govern the organization of the space-economy and the emergence of agglomerations have been understood for a long time. First, the observed spatial configuration of economic activities is the result of a complicated balance of forces that push and pull consumers and firms. These forces may be organized in two main categories: agglomeration (or centripetal) forces and dispersion (or centrifugal) forces. 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" (our translation). Specifically, the location of human activity can be viewed as *the interplay between the need for proximity and a crowding-out effect*: agents benefit from a better proximity to one another or to some place, but face tougher competition in the use of scarce resources such as land and a green environment. As a consequence, production and consumption are locationally interdependent: consumption patterns are determined by the spatial distribution of consumer income, which in turn depends on the location of production, and vice versa.

### 1.4.1 Agglomeration and Increasing Returns

The most natural way to think of increasing returns is to recognize that firms must build a facility or a plant before starting production. This gives rise to overhead and fixed costs, which are typically associated with mass production. In other words, scale economies are *internal* to firms as in standard location theory.

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 transport 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:[5]

> 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. 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

[5] That trade and location theories are two sides of the same coin was explicitly recognized by the Royal Swedish Academy of Sciences in the press release announcing the 1977 Nobel Prize in Economic Sciences: "Ohlin has also demonstrated similarities and differences between interregional (intra-national) and international trade, and the connection between international trade and the location of industries."