PART I

Cities and Agglomeration

CHAPTER 1

The Formation of Economic Agglomerations: Old Problems and New Perspectives

Masahisa Fujita and Jacques-François Thisse

1 Introduction

"Nearly half the world's population and three-quarters of all westerners live in cities" (*The Economist*, July 29, 1995). This raw fact can no longer be given lip service and then put aside. We are therefore led to raise the following fundamental question: *Why do economic activities tend to agglomerate in a small number of places (typically cities)*?

More precisely, we want to try to explain why certain economic activities tend to become established in particular places, and we want to examine the resulting geographical organization of the economy. Intuitively, the equilibrium spatial configuration of economic activities can be viewed as the outcome of a process involving two opposing types of forces, that is, *agglomeration* (or centripetal) *forces* and *dispersion* (or centrifugal) *forces*.[1] This view agrees with some very early work in economic geography. For example, in his *Principes de Géographie humaine*, published 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 in order to benefit from the advantages of the division of labor, but various difficulties restrict the gathering of many individuals. Similarly, Lösch (1940) viewed the economic landscape as the

The authors are grateful to Simon Anderson and Vernon Henderson for helpful discussions during the preparation of this chapter. They also thank Gilles Duranton, Louis-André Gérard-Varet, Yossi Hadar, Jean-Marie Huriot, Yoshitsugu Kanemoto, Xavier Martinez-Giralt, Dominique Peeters, Diego Puga, Tony Smith, and Takatoshi Tabuchi for useful comments. They have also benefited from suggestions and remarks by participants at the trilateral TCER/NBER/CEPR conference on "Economic Agglomeration," the CEPR workshop on "Trade, Location and Technology," the European Summer Symposium in Economic Theory, and seminar audiences at Kyoto University, Université de Bourgogne, and Université Catholique de Louvain. A shorter version of this chapter has been published in the *Journal of Japanese and International Economies*, vol. 10 (1996), pp. 339–78.

[1] The term "agglomeration" is less ambiguous than "concentration," which is used to describe different phenomena. It was introduced in location theory by Weber (1909, ch. 1). Though Weber is known mainly for his work on the location of the firm (Wesolowsky, 1993), his main concern was to explain the formation of industrial clusterings.

outcome of "the interplay of purely economic forces, some working toward concentration and others toward dispersion" (p. 105 of the English translation).

Among the several questions that have been investigated in the literature, the following are central: (1) How are agglomeration and dispersion forces generated? (2) Why do we have cities?[2] (3) Why do various regions and cities specialize in different activities? In order to answer these questions, we must consider a variety of models focusing on different aspects of the economics of cities. Indeed, it would be futile to look for a single model that could explain the economic landscape of economies at different stages of development and in different institutional environments. As mentioned earlier, a model of economic geography must take account of both centripetal and centrifugal forces. The equilibrium spatial configuration of economic activities is then the result of a complicated balance of forces that push and pull consumers and firms until no one can find a better location. As will be seen, the major models that have been developed do reflect such an interplay.

Though convenient at a high level of abstraction, it should be clear that the concept of *economic agglomeration* as used in this chapter refers to a variety of real-world phenomena. For example, one type of agglomeration arises when restaurants, movie theaters, or shops selling similar products are clustered within a single neighborhood of a city. At the other end of the spectrum lies the core–periphery structure corresponding to North–South dualism. For example, Hummels (1995) observed that high-income nations are clustered in small industrial cores in the Northern Hemisphere and that income steadily declines with distance from these cores. Other types of agglomeration can be seen in the existence of strong regional disparities within a given country, in the formation of cities of different sizes, and in the emergence of industrial districts where firms have strong technological and/or informational linkages. This should not come as a surprise, for geographers have long known that scale matters in studying spatial problems. Although we shall consider these different types of spatial clusterings, the main emphasis of this study will be on *city formation*.[3]

In recent years, increasing numbers of economists have become interested in the study of location problems. This is probably best illustrated by the work of Henderson (1988), Lucas (1988), Krugman (1991a,b), and Becker and Murphy (1992), among several others, work that triggered a new flow of interesting contributions in the field. No doubt this increased interest has been fostered by the integration of national economies within trading blocs such as the European

[2] This question bears some resemblance to that raised by Coase concerning the reason for firms to exist, because firms are also formed by clusters of individuals performing different tasks. However, if firms can be viewed as composing the nexus of contracts, cities involve more complex systems of relationships.

[3] We do not necessarily consider cities as being monocentic; see Berry (1993) for a critical appraisal of this model.

Union and the North American Free Trade Agreement and their impact on the development of their regions and cities. As market integration increasingly dissolves economic barriers between nations, national boundaries no longer demarcate the most natural units for analysis (economists still tend to suffer from cartographic illusion). Contrary to widespread opinion, this consideration is not new; it was raised by some scholars at the outset of the discussions that were to lead to the European Union (Giersch, 1949). However, the subject remained neglected for a long time, despite the suggestions made by Ohlin (1933, pt. III), who proposed to unify interregional trade theory and location theory. Nowadays the issue seems even more important, for the continuing growth of trade and especially the development of multinational production systems are casting doubt on the relevance of the concept of national economies. As a result, location theory and studies of international trade are increasingly focusing on economic agglomerations, local specializations, and inter-city trade.

Applications of the new theories of growth are also under scrutiny. The role of cities in economic growth since the second half of the nineteenth century has been emphasized by economic historians (e.g., Hohenberg and Lees, 1985, ch. 6 and 7). Indeed, cities and, more generally, economic agglomerations are considered to be the main institutions in which both technological and social innovations are developed through market and non-market interactions. Furthermore, city specializations can change over time, thus creating a geographically diversified pattern of economic development. For all these reasons, it seems reasonable to say that *growth tends to be localized*, a fact that was recognized by the early theorists of development, such as Myrdal (1957) and Hirschman (1958). This observation has been at the core of many recent empirical contributions that have shed new light on the mechanisms of growth (e.g., Glaeser et al., 1992, 1995; Henderson et al., 1995).[4]

In particular, Feldman and Florida (1994) have observed that in the late twentieth century, *innovations have tended to appear in geographic clusters* in areas where firms and universities oriented toward research and development (R&D) have already become established, and such concentrations of specialized resources reinforce a region's capacity to innovate and to grow. Consequently, the connection between growth and geography becomes even stronger when regional specialization in innovative activities is viewed as the outcome of a

[4] It is worth noting that pre-classical economists have stressed the role of cities in the process of development. See, e.g., Lepetit (1988, ch. 3) for an overview of the main contributions prior to Adam Smith. In particular, they viewed cities not only as a combination of inputs but also as a "mutiplier" that leads to increasing returns in the aggregate. In accord with modern urban economics (discussed later), pre-classical 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 economic geography" and theories of endogenous growth have the same historical roots. There are here several interesting questions that should be explored by historians of economic thought.

combination of specific capabilities and capacities developed in those regions, thus suggesting that the process at work is similar to the one we shall encounter in the formation of agglomerations.

Thus it seems fair to say that the "new economic geography," which can also be termed *geographical economics*, is in many respects more deeply rooted in economic theory than in the traditional theories of location. As we shall see in the course of this study, geographical economics has strong connections with several branches of modern economics, including industrial organization and urban economics, but also with the new theories of international trade and theories of economic growth and development. This suggests that this field has considerable potential for further development and that cross-fertilization can be expected (e.g., Ioannides, 1994; Martin and Ottaviano, 1996; Palivos and Wang, 1996; Walz, 1996). These developments have generated a large flow of empirical studies that have used the modern tools of econometrics, thus leading to more firmly grounded conclusions.

As in any economic field, several lines of research have been and are being explored in geographical economics. The earliest line was initiated by von Thünen (1826), who sought to explain the pattern of agricultural activities surrounding many cities in pre-industrial Germany.[5] More generally, von Thünen's theory has proved to be very useful in studying land use in situations in which economic activities are perfectly divisible (Mills, 1970). In fact, the principles underlying his model are so general that von Thünen can be considered the founder of marginalism (Nerlove and Sadka, 1991). Despite the fact that we now recognize his monumental contributions to economic thought, von Thünen's ideas languished for more than a century without attracting widespread attention. (Note that the same holds for other contributions to location theory, despite the efforts of some scholars to make that literature accessible to a large audience of economists at its very beginning; see, e.g., the survey offered by Krzyzanowski, 1927). Yet, following a suggestion made by Isard (1956, ch. 8), Alonso (1964) succeeded in extending von Thünen's central concept of bid-rent curves to an urban context in which a marketplace was replaced by an employment center (the "central business district"). Since that time, urban economics has advanced rapidly. Furthermore, as observed by Samuelson (1983), the von Thünen model also contains the basic ideas of comparative advantage on which other economists have built the neoclassical theory of international trade. The reason for such a broad range of applications lies in the fact that the model is compatible with the competitive paradigm, because production takes place under constant returns to scale.

However, the von Thünen model has several limitations. Indeed, the following question suggests itself: Why is there a unique city in von Thünen's isolated

[5] Note that the von Thünen model has been reformulated in mathematical terms by Launhardt (1885, ch. 30).

state? Or why a unique central business district in most urban economic models? Though such a center may emerge under constant returns when space is heterogeneous (Beckmann and Puu, 1985), this is more likely to occur when increasing returns are at work in the formation of trading places or in the production of some goods; in other words, one must appeal to something that is not in the von Thünen model to understand what is going on.

Conceding the point, Lösch (1940) argued that scale economies in production, as well as in transportation costs, are essential for understanding the formation of economic space. He then proceeded to construct a model of monopolistic competition in the manner of Hotelling and Kaldor as an alternative to von Thünen's model.[6] Lösch's model is still used as a reference in "classical" economic geography, but it differs from the Dixit-Stiglitz model employed in the "new" economic geography discussed later in Section 3.1. In the same spirit, Koopmans (1957, p. 157) made it clear that scale economies are essential in the creation of urban agglomerations: "without recognizing indivisibilities – in the human person, in residences, plants, equipment and in transportation – urban location problems down to the smallest village cannot be understood."

The assumption of nonincreasing returns indeed has dramatic implications for geographical economics that help us understand why so many economists have been tempted to put space aside. Under nonincreasing returns and a uniform distribution of resources, the economy would reduce to a Robinson Crusoe type, where each individual would produce only for his or her own consumption (backyard capitalism). Mills (1972, p. 4) provided a neat description of such a world without cities:

> land would be the same everywhere and 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.

Each location could thus be a base for an autarkic economy, where goods would be produced on an arbitrarily small scale, except possibly (as in the neoclassical theory of international trade) that trade might occur if the geographical distribution of resources was nonuniform. Although pertinent (Courant and Deardoff, 1992; Kim, 1995), an unequal distribution of resources seems insufficient to serve as the only explanation for specialization and trade (Ciccone and Hall, 1996). Furthermore, when capital and labor can move freely, the neoclassical model of trade does not allow for prediction of the sizes of regions

[6] See Beckmann (1972) for a modern presentation of this model.

when natural resources are uniformly distributed. Accordingly, nothing can be said about the location of production activities within this model. We can therefore safely conclude that *increasing returns to scale are essential for explaining the geographical distribution of economic activities*.[7] However, when indivisibilities are explicitly introduced, the nonexistence of a competitive equilibrium in a spatial economy is common, as shown by Koopmans and Beckmann (1957) and Starrett (1978).[8]

Furthermore, as noticed by Drèze and Hagen (1978) in a somewhat different context, scale economies in production have another far-reaching implication: The number of marketplaces open at a competitive equilibrium is likely to be suboptimal. Or, to use a different terminology, *spatial markets typically are incomplete*, so that an equilibrium allocation is, in general, not Pareto-optimal. More precisely, there are various levels of Pareto optimality corresponding to different environments, as in club theory (Scotchmer, 1994).

A combined consideration of space and economies of scale has one further implication that turns out to be even more fundamental for economic theory. If production involves increasing returns, a finite economy can accommodate only a finite number of firms that are imperfect competitors. Treading in Hotelling's footsteps, Kaldor (1935) argued that space gives this competition a particular form. Because consumers will buy from the firm with the lowest "full price," defined as the posted price plus the transport cost, each firm competes directly with only a few neighboring firms, regardless of the total number of firms in the industry. The very nature of the process of spatial competition is therefore oligopolistic and should be studied within a framework of interactive decision-making. That was one of the central messages conveyed by Hotelling (1929), but it was ignored until economists became fully aware of the power of game theory for studying competition in modern market economies (see Gabszewicz and Thisse, 1986, for a more detailed discussion). Following the outburst of industrial organization that began 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.[9]

[7] This statement, which goes back at least to Lösch (1940, ch. 9), has been rediscovered periodically. For this reason, it can be referred to as the "folk theorem of geographical economics" (see Scotchmer and Thisse, 1992, for a more detailed discussion). In the same vein, planning models of location developed in operations research have also emphasized the trade-off between fixed production costs and transportation costs; see Manne (1964) and Stollsteimer (1963) for early contributions. A recent survey of those models has been presented in Labbé et al. (1995).

[8] The nonexistence of a competitive equilibrium in the presence of indivisibilities is of course related to the possibility of observing duality gaps in integer programming, that is, the primal and the dual take different values at the optimal solution.

[9] Simultaneously, new developments in local public finance have led some to question the relevance of the Samuelsonian paradigm of (pure) public good. There are interesting analogies and contrasts between these two lines of research (Scotchmer and Thisse, 1992).

Conversely, "space" is often used in various economic areas as a label for describing nongeographical characteristics along which economic agents are heterogeneous. In particular, such an approach has been followed in many models of industrial organization.[10]

Despite its factual and policy relevance, the question of why a hierarchical system of cities emerges remains open. In particular, it is a well-established fact that cities tend to be distributed according to some specific relationship relating their size and their rank in the urban system (what is called the rank-size rule). The first attempts to build a spatial theory of the urban hierarchy date back at least to the German geographer Christaller (1933), who pioneered "central place theory," based on the clustering of marketplaces for different economic goods and services.[11] Though the theory proposed by Christaller and developed by Lösch has served as a cornerstone in classical economic geography, as described by Mulligan (1984) in a nice overview, it is fair to say that the microeconomic underpinnings of central place theory are still to be developed. See Henderson (1972) for an early critical, economic evaluation of this theory and Hohenberg and Lees (1985, ch. 2) for an appraisal from the historical perspective.

The topic is difficult because it involves various types of nonconvexities that are even more complex to deal with than are increasing returns in production. For example, a consumer organizes his shopping itinerary so as to minimize the total cost of purchases, including transport costs. This problem is extremely complex: Determining the optimal geographical structure of purchases requires solving a particularly difficult combinatorial problem, and finding an equilibrium becomes very problematic (Bacon, 1984). In the same vein, often there are considerable scale economies in carrying the goods bought by a consumer when shopping. In the extreme, consumers' outlays on transportation can be considered as independent of the quantities purchased. These nonconvexities affect demand functions in complex ways that have not been fully investigated. This is just one example of the many difficulties one encounters in attempting to construct a general spatial model that can account for cities of different sizes trading different commodities. It is therefore no surprise that we still lack such a model, because it is well known that economic theory has serious problems in dealing with nonconvexities. Yet this turns out to be a real embarrassment, because the rank-size rule is one of the most robust statistical relationships known so far in economics (Krugman, 1995, ch. 2).

A major centripetal force can be found in the existence of externalities (later discussion will clarify what we mean by "externality"), in that a geographical

[10] Examples include the supply of differentiated products (Ireland, 1987), the various forms of price discrimination (Phlips, 1983), and the competition between political parties (Enelow and Hinich, 1984). Other applications, in particular in labor economics, are possible.

[11] Note that this problem bears some resemblance to that of the firm size distribution studied in the "old" industrial-organization literature.

concentration of economic activities can be viewed as the outcome of a snowball effect.[12] Specifically, more and more agents want to agglomerate because of the various factors that will allow for greater diversity and higher degrees of specialization in the production processes, leading to a wider array of products available for consumption. The setting up of new firms in such regions gives rise to new incentives for workers to migrate there because they can expect better job matching and therefore higher wages. This in turn makes the place more attractive to firms, which may expect to find the types of workers and services they need, as well as new outlets for their products. Hence, both types of agents benefit from being together. This process has been well described by Marshall (1890, 1920, p. 225):

> 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 neighborhood to one another. . . . A localized industry gains a great advantage from the fact that it offers a constant market for skill. . . . Employers are apt to resort to any place where they are likely to find a good choice of workers with the special skill which they require; while men seeking employment naturally go to places where there are many employers who need such skills as theirs and where therefore it is likely [they will] find a good market.

More generally, the "Marshallian externalities" arise because of (1) mass production (the so-called internal economies that are similar to the scale economies mentioned earlier), (2) the 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, (3) the availability of specialized input services, and (4) the existence of modern infrastructures. Not surprisingly, Marshallian externalities provide the engine for economic development in the new growth theories.[13]

Building on Weber (1909, ch. 5), Hoover (1936, ch. 6) has proposed what has become the now-standard classification of agglomeration economies (see also Isard, 1956, ch. 8): *scale economies* within a firm, depending upon the size of the firm's scale of production at one point; *localization economies* for all firms in one industry at one point, depending upon the total output of the industry at that location;[14] *urbanization economies* for all firms in various industries at one

[12] This phenomenon is similar to that encountered in studies of network externalities (David and Greenstein, 1990). 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 (Belleflamme, 1998). Note also that the issue of standardization bears some resemblance to that of agglomeration (Arthur, 1994, ch. 2 and 4). Finally, the stratification of a population can be described by a similar cumulative process (Bénabou, 1996a).

[13] They are also at the heart of some early contributions to studies of economic development (see Section 3).

[14] See Chipman (1970) for an early formal analysis of these externalities developed in a nonspatial model.

point, depending on the overall level of activity at that location. Scale economies correspond to Marshallian externalities of type (1); localization economies refer to Marshallian externalities of types (2) and (3); urbanization economies would cover the Marshallian externalities of types (2), (3), and (4), since they typically depend on the presence of public infrastructures and on the agglomeration size (which in turn depends on the division of labor within the city). This classification has been used extensively in empirical studies, as surveyed by Henderson (1988, ch. 5).

The advantages of proximity for production have their counterpart on the consumption side. For example, cities typically are associated with a wide range of products and a large spectrum of public services, so that consumers can reach higher utility levels and therefore will have stronger incentives to migrate toward cities. Furthermore, the propensity to interact with others, the desire of man for man, is a fundamental human attribute, as is the pleasure of discussing and exchanging ideas with others. Distance is an impediment to such interactions, thus making cities the ideal institution for the development of social contacts corresponding to various kinds of externalities (Fischer, 1982, ch. 2 and 3). Along the same line, Akerlof (1997) has argued that the inner city is the basis for the development of social externalities (e.g., conformity and status-seeking) that govern the behaviors of particular groups of agents. For example, social capital arising across individuals living within the same city (or neighborhood) has been explored by Bénabou (1993, 1996a), who has shown its importance for urban development.

Before describing the content of this chapter, we want to clarify the following issue. For many years, the concept of *externality* (also called *external effect*) has been used to describe a great variety of situations. Following Scitovsky (1954), it has been customary to consider two categories: *technological externalities* (such as spillovers) and *pecuniary externalities*. The former deals with the effects of non-market interactions that are realized through processes directly affecting the utility of an individual or the production function of a firm. By contrast, the latter refers to the benefits of economic interactions that take place through the usual market mechanisms via the mediation of prices. For obvious reasons, Marshall was not aware of this distinction, and his externalities turn out to be mixtures of technological and pecuniary externalities. As a consequence, each type of externality may lead to the spatial agglomeration of economic activities.

In order to understand how an agglomeration occurs when Marshallian externalities are present, it is useful to divide human activities into two categories: *production* and *creation*. Roughly speaking, one can say that production encompasses the routine ways of processing or assembling things (such as the preparation of a dinner or the working of an assembly line). For an agglomeration of firms and households to be based on this type of production activity, the presence of pecuniary externalities is crucial.