Journal of Economic Literature 2019, 57(3), 575–643
https://doi.org/10.1257/jel.20181414

What Can Be Learned from Spatial Economics?[†]

Stef Proost and Jacques-François Thisse[*]

Spatial economics aims to explain why there are peaks and troughs in the spatial distribution of wealth and people, from the international and regional to the urban and local. The main task is to identify the microeconomic underpinnings of centripetal forces, which lead to the concentration of economic activities, and centrifugal forces, which bring about the dispersion of economic activities at the regional and urban levels. Transportation matters at both scales, but in a different way. The emphasis is on the interregional flows of goods and passenger trips at the regional level and on individual commuting at the urban level. (*JEL* F12, L13, R12, R23, R30, R40)

1. *Introduction*

Spatial economics deals with bringing *location*, *transport*, and *land* into economics. These three concepts are closely intertwined and gathered under the category R of the JEL Classification System. We show how spatial economics addresses many issues, including the following: Why are economic activities unevenly distributed across space at different spatial scales? Does the steady drop in transport costs since the mid-nineteenth century, compounded by the near-disappearance of communication costs, imply that distance and location have disappeared from economic life? Why do lasting and sizable regional disparities exist in many countries? Why do firms locate in areas where labor and land are expensive? Does building interregional transport infrastructures help to reduce inequality across space? Why do cities exist and why do they differ in size? Why are workers better paid and housing more expensive in large than in small cities? Are workers sorted by skills across cities? Is road pricing the ideal instrument to tackle traffic congestion? We discuss various approaches, ranging from classical location theory to quantitative spatial models.

Spatial economics is both at the core and periphery of economics. It is at the core in that economic growth has always been, and still is, geographically localized and uneven, while economic historians have convincingly argued that cities have played a fundamental

[*] Proost: Department of Economics, KULeuven. Thisse: CORE-UCLouvain, HSE University and CEPR. We are very grateful to Steven Durlauf and three referees for many comments and suggestions. We also thank Jan Brueckner, Miren Lafourcade, Sun Kyoung Lee, Robin Lindsey, Mario Polèse, Stephen Redding, and Kenneth Small for insightful remarks. The second author acknowledges the financial support of the Russian Science Foundation under grant 18-18-00253.

[†] Go to https://doi.org/10.1257/jel.20181414 to visit the article page and view author disclosure statement(s).

role in the process of economic and social development (Bairoch 1988, Hohenberg and Lee 1985). Therefore, space and its parent concepts—such as location, transport, and land—should be an important component of economists' toolbox. However, spatial economics is fraught with many of the difficulties encountered in economic theory (externalities, increasing returns, and imperfect competition). Therefore, it hardly comes as a shock that spatial economics and its constituent subfields—such as regional, urban and transportation economics—were, and are still, at the periphery as they are poorly represented in textbooks and graduate programs. Yet, the empirical evidence is compelling.

By using nighttime images, Florida (2017) finds that the largest 681 metropolitan areas distributed over the globe house about 24 percent of the world's population and produce 60 percent of the global output. In the United States, 20 metropolitan areas produce about 50 percent of the GDP but are home to approximately 30 percent of the population. By contrast, the 28 capital regions of the European Union accounted for 23 percent of its GDP. Greater Paris, which represents 2 percent of the area of continental France and 19 percent of its population, produces more than 30 percent of its GDP. Equally surprising, distance still matters in international trade. One of the most robust empirical facts in economics is the gravity law linking bilateral trade flows to the GDP of countries and the distance between them. Head and Mayer (2014) summarize the state of the art as follows: "Holding constant the product of two countries' sizes, their bilateral trade will, on average, be inversely proportional to the distance between them." These facts and many others run against the common belief that we live in a world where the playground has been leveled. Despite spectacular drops in communication and transport costs, distance and location have not disappeared from economic life. New forces, hitherto outweighed by natural factors, are shaping an economic landscape that, with its many barriers and large inequalities, is anything but flat.

Spatial economics deals with other important issues. Housing and transport, which are quintessentially space-related commodities, rank first and second in household expenditure. In the United States, the average share of household expenditures on housing is 24 percent, while it is 33 percent in France. Spending on transport amounts to 17 percent in the United States and 13.5 percent in France. Per year, the opportunity cost of the time spent in commuting comes to three to six weeks of work for a typical New Yorker and, on average, to four weeks of work for a resident of Greater Paris. These numbers do not account for the fact that commuting seems to be one of the most unpleasant activities in individuals' daily life (Kahneman et al. 2004).

Space is the substratum of human affairs, but it is also a consumption and production good in the form of *land*. The worldwide supply of land is perfectly inelastic but vastly exceeds the demand for it. Therefore, putting aside the agricultural land rent, the price of land should be zero. Yet, housing costs may be very high and vary enormously with the size and composition of cities for reasons that do not depend on the quality of the housing structure. According to Albouy, Ehrlich, and Shin (2018), in 2006 (i.e., before the subprime mortgage crisis), the total value of urban land in the United States exceeded twice the American GDP. Therefore, the price of land in cities must reflect the scarcity of "something" that differs from land per se.

Locations are linked through various types of flows: of goods, people (commuters and passengers), factors of production (capital and labor), and information. Therefore, one could expect trade theory to be the economic field that paid the most attention to the spatial dimension. Yet, for a long time, trade theory focused on an extremely strange geography with countries that are close enough for the cost of shipping goods internationally to be zero, but distant enough for no workers or capital owners to find their way from one country to the other (Leamer 2007). This research strategy is especially surprising since Ohlin challenged the common wisdom long ago: "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" (1968 [1933], p. 97). Eventually, the gravity law led trade economists to recognize, if belatedly, that free trade does not mean costless trade.

Economists and regional scientists have developed different theories and models to account for the existence of a variety of spatial clusters. However, the differences among theories and models are less pronounced than they might seem. One of the main thrusts of this survey is to show that a few basic principles may be used to understand the reasons for a great number of spatial clusters. In what follows, we discuss regional, urban, and transportation economics. What distinguishes regional from urban economics is not a simple issue. We contend that the main difference lies in the spatial unit of reference.

Regional economics focuses on the mobility of goods and production factors within a large, economically integrated space (e.g., a nation or trade bloc). Because it focuses on issues arising on a global scale where spillovers are likely to be absent, regional economics relies on the combination of increasing returns and imperfect competition, while trading goods across regions is costly. Since market prices do not accurately reflect the social value of individual decisions, pecuniary externalities matter. Until recently, regional economics has neglected land, and thus may be viewed as spatial economics *without* land. Urban economics studies the formation of cities, their spatial structure, and their social composition. As cities differ in size and specialization, they form what is called the urban system. Urban economics is spatial economics *with* land because land use is critical in the working of cities. Urban economics abides to technological externalities through market size effects and knowledge spillovers.

Transportation economics spans both regional and urban economics, but each in a different way. In the regional context, transportation economics studies the interregional and international freight trips of inputs and outputs, as well as passenger trips (whether for business or leisure). In the urban context, the emphasis is mainly on commuting by various means of transportation. Most of transportation economics takes the location of firms and households as given. This is where the link is missing with regional and urban economics, where locations are endogenous. Thus, there is a need for more integration between, on the one hand, regional or urban and, on the other, transportation economics. The literature in spatial economics has paid too little attention to what has been accomplished in transportation economics (and vice versa).

The paper is organized as follows. Section 2 discusses the trade-off between increasing returns and transport costs, which shapes the space-economy at different scales. In addition, we explore a question that has generated endless discussions: how best to describe the process of competition across space. To answer this question, we borrow tools from general equilibrium theory and industrial organization. Section 3 focuses on the so-called "regional question," that is, the existence of sizable and lasting disparities in GDP per capita within nations such as the Italian regional divide, or regional trade blocks characterized by strong spatial inequality like the European Union. Speculation about regional disparities has never been in short supply. However, no one before Krugman (1991) had proposed a full-fledged general equilibrium model showing how these disparities can arise at a stable equilibrium in a setting involving market forces alone. Once set in motion, these forces generate “snowball effects” that are self-reinforcing.

We follow in Krugman’s footsteps and adopt the “new economic geography” approach to regional economics. What distinguishes this field from trade theory is the mobility of firms and households. Admittedly, attention to new economic geography has waned during the last decade. However, it seems to have regained interest thanks to the application of new analytical tools, while its main ideas have been taken on board in quantitative spatial models. In addition, since the key parameter of new economic geography is the level of transport costs, this approach is a natural candidate for studying the effects of trade and transport policies on the location of economic activity.

In section 4, we move on to studying cities, which are the most extreme form of spatial inequality. Urban economics, in the wake of its founding fathers (Alonso 1964, Mills 1967, Muth 1969), focused for a long time on the monocentric city model in which jobs are supposed to be concentrated in the city’s central business district. One of the great merits of this model is its emphasis on the importance of non-tradables produced within the city for residents, whereas regional economics remains in the tradition of trade theory by focusing on tradables. Ever since the early 1970s, interest in urban economics has advanced rapidly and shows no sign of abating. The reason for this success is probably that the monocentric city model is based on the perfectly competitive paradigm. More recently, the interest has shifted to the reasons explaining the agglomeration of firms and households in a relatively small number of cities. In this context, cities are not just the containers of economic activities, but rather key players in the social fabric. Agglomeration economies explain why cities exist and why city size matters. Congestion is the main friction in cities, and transportation economics offers some promising insights for easing this friction. In section 5, we discuss different families of models that could be used to develop a synthesis of regional and urban economics, including urban systems and quantitative spatial modeling. Section 6 concludes with suggestions for future research.

Even though ongoing research focuses mainly on the empirics of agglomeration, our emphasis is primarily on theory. The fifth volume of the *Handbook of Regional and Urban Economics* features several insightful chapters discussing the recent empirical literature (Duranton, Henderson, and Strange 2015). By contrast, many ideas and results obtained in *spatial economic theory* have been forgotten or are being rediscovered under different guises. Furthermore, they remain fairly disparate and in need of a synthesis. In this survey, we aim to show that the state of the art is sufficiently advanced to sketch a unified field that should attract more attention in the economics profession. Yet, doing so does not mean that we may leave empirical works aside. Reflecting our idiosyncrasies, we have chosen those that are directly related to the bodies of theory we discuss.

2. *What Are the Specificities of Spatial Economics?*

In this section, we first expand on the spatial concentration of activities by going beyond purely geographical factors. Next, we introduce spatial economics’ fundamental trade-off between increasing returns to scale and transport costs. Somewhat unfortunately, this trade-off does not fit into the competitive general equilibrium model. This is why less standard assumptions, such as spatial inhomogeneities, spatial externalities, and imperfect competition, have been important for the development of the field. We conclude with a discussion of the strategic locational choices made by firms, a topic that underpins spatial equilibrium analysis.

2.1 *Does Geography Matter?*

According to Diamond (1997), spatial differences between edible plants (with abundant nutrients) and wild animals (capable of being domesticated to help man in his agricultural and transport activities) explain why only a few regions have become independent centers of food production. Though relevant for explaining the emergence of civilization in a few areas, we must go further to understand why, in the wake of the Industrial Revolution, trade and cities have grown so rapidly.

It is true that a number of natural features differentiate locations and they may have long-run implications for the organization of the space-economy (Gallup, Sachs, and Mellinger 1999; Bosker and Buringh 2017). The fact that the two biggest American cities, New York and Los Angeles, are also the two biggest ports of the United States is probably not an accident. However, Moscow and Paris, which rank first and third among the top European cities, have no serious comparative advantages in terms of access to the sea nor any other great natural amenities. In this case, one must consider historical and political factors, rather than geographical ones, to explain the importance of these two cities. In this survey we have chosen to follow the founding fathers of spatial economics, such as Lösch (1940) and Hoover (1948), in their methodological choice. For these authors (and for us), the issue that should rank first on the research agenda is to explain the existence of different types of economic agglomerations through the interplay of economic forces—some working toward concentration, and others toward dispersion—in spaces that are otherwise undifferentiated by natural or political factors. Given this methodological choice, we are left with some fundamental questions: What are the macro-spatial consequences of a myriad of micro-decisions made by economic agents? What are the main economic forces that push toward the agglomeration or the dispersion of firms and households? How do spatial frictions, such as transport and commuting costs, affect the economic landscape?

This choice does not reflect any prejudice on our part. We do believe that natural factors matter for the organization of the space-economy. In a recent and fascinating paper, Henderson et al. (2018) show that 35 percent of within-country variation in night light, used as a proxy for the level of economic activity, is associated with physical geography. The remaining 65 percent are the raison d'être of spatial economics. Moreover, we also believe that natural factors do not explain why huge metropolitan areas such as Tokyo, New York, or Shanghai exist. Geography is more useful in understanding *where* cities are rather than *why* cities exist, even though both probably interact in some way. Thus, when the aim is to explain the actual location of economic activity, natural factors must be taken into consideration. According to Allen and Arkolakis (2014), 20 percent of the spatial variation in income across the United States in 2000 can be explained by geography alone.

Geography also matters in a more subtle way: the choice of the *spatial scale*. Quite often, economists use different, yet equally unclear, words such as locations, regions, or places interchangeably without being aware that they often correspond to different spatial units. In so doing, they run the risk of drawing implications that are valid at a certain spatial scale but not at another, because the agglomeration or dispersion forces at work at the local level are not necessarily the same as those acting at the global level.

2.2 *The Fundamental Trade-off of Spatial Economics*

Our aim in this section is to explain *the fundamental trade-off between increasing returns and transport costs*, a trade-off that is valid at all spatial scales (Fujita and Thisse 2013). The intuition behind this trade-off is easy to grasp. In the absence of increasing returns, firms would be able to spread their production over an arbitrarily large number of locations without any efficiency loss, while bringing transport costs down to zero. In this case, the economy boils down to *backyard capitalism*, that is, a world in which Friday is no help to Robinson Crusoe. If transport costs were nil, with increasing returns, firms would concentrate their production in a few giant plants to benefit from the highest possible level of efficiency. Such counterfactuals confirm the relevance and importance of the fundamental trade-off of spatial economics. It has been rediscovered several times and goes back at least to Lösch (1940). We now briefly discuss the two main ideas that stand behind this trade-off and serve as the ground rules for this survey.

2.2.1 *Scale Economies Matter for the Location of Economic Activity*

The most natural way to think of increasing returns is when a plant needs a minimum capacity to operate. This gives rise to overhead and fixed costs, which are typically associated with mass production. In this case, scale economies are *internal* to firms. Similarly, local public goods are often provided in the form of a facility designed to supply collective services to consumers (e.g., fire stations or public hospitals), or as a road or subway network. Increasing returns may also materialize under a very different form, where they are *external* to firms but specific to the local environment in which firms operate. Their concrete manifestation can vary considerably from one case to another, but the basic idea is the same: each firm benefits from the presence of other firms.

The presence of increasing returns has a major implication for the spatial organization of the economy: not everything can be produced everywhere. It is in this sense that location matters. Though a large number of activities became "footloose," many large areas in many countries account for no or little economic activity. Indeed, one should not infer from the low value of (monetary and nonmonetary) transport costs that location matters less. It is quite the opposite. In the presence of increasing returns, low transport costs make firms more sensitive to minor differences between locations. All of this is seemingly a paradox: the inexpensive shipping of goods makes competition tougher, and thus firms care more about small advantages than they do in a world where they are protected by the barriers of high transport costs.

2.2.2 *Moving Goods, People, and Information Remains Costly*

A high number of personal services are not tradable, i.e., they must be consumed where they are produced. In addition, what has been built, e.g., roads, subways, and housing, cannot be moved. In those cases, transport costs may be considered as infinite.

Goods.—The great many estimations of the gravity equation show that distance remains a strong impediment to the spatial exchange of goods (Anderson 2011, Head and Mayer 2014). Anderson and van Wincoop (2004) estimate that the average cost of delivering a good from the point of manufacture to the destination represents 170 percent of the producer price, which is quite high. However, this number is somewhat inflated as it includes retail costs. This is not an isolated finding. Even within-country trade tends to be very much localized. Using data on shipments by individual US manufacturing plants, Hillberry and Hummels (2008) find that the number of commodities shipped and the number of plant–destination pairs fall dramatically with distance. The value of shipments within a 4-mile radius of the plant is 3 times higher than those beyond this radius.

People.—To the best of our knowledge, there is no integrative work on migration comparable to that in the trade literature on the gravity equation, though several papers suggest that frictions exist that would make migration gravitational in nature (e.g., Kennan and Walker 2011). As the mobility of people is governed by a broad range of economic, intangible, time-persistent, and idiosyncratic factors, it is more difficult to study than trade flows. In a classical paper devoted to the evolution of the US states over the 1950–90 period, Blanchard and Katz (1992) find that "the dominant adjustment mechanism is labor mobility, rather than job creation or job migration. Labor mobility, in turn, appears to be primarily a response to changes in unemployment, rather than in consumption wages." On the other hand, using the British Household Panel Survey from 1991 to 2009, Bosquet and Overman (2019) observe that 43.7 percent of workers worked only in the area in which they were born.

That many people continue living in deprived areas should be evidence that migration costs are of first-order importance. Due to family bonds, social networks, tacit information, and language and cultural differences, people are heterogeneous in their attitudes toward migration. When individuals are identical and perfectly mobile, the spatial equilibrium condition implies the equalization of utility across space (see Alonso 1964, Henderson 1974, Roback 1982 for the pioneering contributions). We recognize that this assumption vastly simplifies the analysis. However, it is far from being innocuous. For example, the perfect mobility of workers implies that the local labor supply is infinitely elastic. By contrast, when workers are imperfectly mobile, the local labor supply has a finite elasticity. The empirical and theoretical implications of assuming perfectly mobile individuals must be interpreted with care because mobility costs affect the spatial arbitraging between amenities, housing, and wages. Describing migration choices by means of discrete choice models is a possible way out.

Information.—It is perhaps more surprising that informational frictions remain substantial in certain activities. Whereas the media steadily stress the globalization of finance, the empirical evidence highlights the importance of distance between lenders and borrowers in explaining the design of loan contracts (Hollander and Verriest 2016). Likewise, there must be compelling reasons for businesspeople to meet in person despite the high opportunity cost associated with traveling. Here, Giroud (2013) documents that the opening of new airline links, reducing the travel time between headquarters and plants, has generated an increase of 7 percent in plants' productivity in multiunit US firms. Campante and Yanagizawa-Drott (2018) show that better passenger connections through longer-range flights (larger than six thousand miles) generate more business links and capital flows with the major hubs in Europe and improve local economic activity in the vicinity of these hubs.

The existence of transfer costs has a major implication for the spatial organization of the economy: these costs imply that what is near to us matters more than what is far. In the presence of indivisibilities, transfer costs are unavoidable because there is a need for proximity to sites that are spatially separated. Consequently, firms and consumers operate within a system of push-and-pull forces whose balance is their most preferred locations.

The above discussion leads us to formulate what we see as the fundamental trade-off in spatial economics:

> *The spatial distribution of activities is the outcome of a trade-off between different types of scale economies and costs that are generated by the transfer of people, goods, and information.*

This trade-off—the unifying idea of this survey—is key in the balance between agglomeration and dispersion forces and is valid on all spatial scales (cities, regions, countries, and continents). Pomeranz (2000) confirms the historical relevance of this trade-off when he observes that the geographical distribution of resources played a key role in the success of England and the failure of China in the Industrial Revolution. Distances between coal deposits in Northwest China and the Yangtze Delta, which hosted a large number of skilled workers, were very large. Thus, transportation costs were much too high to ship coal to the artisans who could have been able to combine it with new technologies. By contrast, distances were considerably shorter in England. Coal, skills, and demand were close to each other at a time where shipping coal was still very expensive and the transmission of knowledge mainly oral. This allowed the corresponding English sites to benefit from scale economies in the production of steel and metallic items.

2.3 *Space and General Equilibrium*

2.3.1 *The Spatial Impossibility Theorem*

The history of the relationship between spatial economics and general equilibrium theory is both complex and obscure. It is complex as it is fraught with difficulties—such as increasing returns and imperfect competition—that have been put aside for simplicity. It is obscure because the several attempts, made over the last fifty years, to clarify this relationship have just muddled the debate with confusing answers. Yet, as illustrated by the following example, the intuitive argument is beautifully simple (Koopmans and Beckmann 1957).

Consider an economy with one worker and one producer who are price takers and with two locations, A and B, each of which is endowed with one unit of land. In equilibrium, the producer is located at A; he produces one unit of the consumption good by using one unit of land and one unit of labor. The worker is located at B; she consumes one unit of the consumption good and one unit of land and supplies one unit of labor. Assume first that the equilibrium land rent at A is higher than or equal to that in B ($R_A^* \geq R_B^*$). Let p_A^* (p_B^*) be the price of the consumption good at A (B) and w_B^* (w_A^*) the wage paid to the worker at B (A). Let $t > 0$ be the unit cost of shipping the consumption good between A and B and $\tau > 0$ the worker's commuting cost between A and B. The no-arbitrage condition implies that the price (wage) gap is just equal to the cost of moving the good (the worker) between A and B; therefore, it must be that $p_B^* = p_A^* + t$ ($w_A^* = w_B^* + \tau$). If the producer were located at B, he would earn a higher profit because he bears a lower production cost ($w_B^* + R_B^* < w_A^* + R_A^*$) and earns a higher revenue ($p_B^* > p_A^*$) by saving transport costs without paying a higher rent. Therefore, there is no competitive equilibrium such that $R_A^* \geq R_B^*$. Assume now that the competitive equilibrium is such that $R_A^* < R_B^*$. In this case, the worker would be better off at A because she earns a higher labor income while paying a lower rent. Since the level of equilibrium prices is unspecified, there is no competitive equilibrium such that neither the producer nor the worker wants to change place. Thus, for a competitive equilibrium to exist, the worker and the producer should be located together, which is impossible because there is not enough land in A and B. Starrett (1978) gives a general proof.

THE SPATIAL IMPOSSIBILITY THEOREM: *Consider an economy with a finite number of locations. If consumers' preferences are the same across locations, firms' production functions are the same across locations, and transport is costly, then there exists no competitive equilibrium involving trade across locations.*

Thus, if a competitive equilibrium exists, the spatial impossibility theorem implies that each location operates as an autarky. If there is no efficiency reason for firms to distinguish between locations and if activities are divisible, each one can operate at an arbitrarily small level and transport costs are reduced to zero. By contrast, when there are indivisibilities in production, some goods must be traded across locations. In this case, what makes a location desirable depends on the locations chosen by the other agents. By implication, the price system must perform two different jobs simultaneously: (i) support trade between locations (while clearing markets in each location), and (ii) prevent firms and consumers from relocating. The spatial impossibility theorem says that it is impossible to hit two birds with one stone. The equilibrium prices supporting trade carry the wrong signals from the viewpoint of locational stability. As a consequence, the price system does not convey all the information needed by an agent. Note the importance of the following indivisibility for the above argument: the land that the worker (the producer) lives on is concentrated in a single location. In other words, each agent has an "address" in space.

2.3.2 *Spatial Inhomogeneities*

There are different strategies to obviate the consequences of the spatial impossibility theorem. The first consists of introducing exogenous spatial inhomogeneities, such as in Ricardian models of trade where firms produce under perfect competition and constant returns at a total factor productivity drawn randomly, which reflects regions' technological comparative advantage (Eaton and Kortum 2002). Admittedly, appealing to spatial inhomogeneities may be a fruitful modeling strategy. For example, assuming the existence of a central business district where jobs are concentrated enables one to study the structure of the monocentric city while using general competitive analysis (see section 4.2). Yet, the question of why this center exists is omitted. Assuming that some locations are endowed with attributes that make firms more productive than elsewhere is tantamount to assuming the answer to a question we want to explain: Why are there differences in productivity across space? Even though there might be a few suitable places to host particular production activities, relying solely on comparative advantage to explain the existence of large urban agglomerations and sizable trade flows amounts to playing Hamlet without the prince.

To fix ideas, consider the above example where it is now assumed that location A is endowed with a strong productivity advantage and location B with a strong amenity advantage. In this case, a competitive equilibrium involving trade exists because the firm and the worker have strongly diverging preferences for location attributes. This type of assumption and result conflicts with our aim to identify conditions under which firms and workers choose to locate together. By contrast, we acknowledge that accounting for locational and natural advantages is needed in empirical research (Diamond 2016, Lee and Lin 2018). However, care is needed in order to avoid using a deus ex machina.

The other two solutions that allow one to escape the spatial impossibility problem became the basis for much of modern developments in spatial economics. In regional economics, which regained interest in the 1990s under the heading of "new economic geography" (Krugman 1991), one calls upon monopolistic competition and internal economies of scale. In urban economics, spatial externalities regained interest with the work of Glaeser et al. (1992) and of Henderson, Kuncoro, and Turner (1995). However, before proceeding, we want to underscore the idea, developed in the spatial and industrial organization literature, that space would give rise to a special type of oligopolistic competition. The detour is important by itself, but it also helps in understanding the modeling choices made in the subsequent sections.

2.4 *Spatial Competition*

In his review of Chamberlin's (1933) book, Kaldor (1935) argued that consumers' dispersion molds competition across space in a very specific way: whatever the total number of firms in the industry is, each one competes more vigorously with its immediate neighbors than with more distant firms. Such market description, which came to be known as *spatial competition*, departs radically from the perfectly competitive setting. The argument goes as follows. Since consumers are spatially dispersed, they differ in their access to the same firm. Consequently, a consumer buys from the firm with the lowest full price, that is, the posted price plus the travel cost to the corresponding firm. This in turn implies that every firm has some monopoly power over the consumers situated in their vicinity, which enables firms to choose their own prices in a strategic environment.

The earliest analyses of this problem are by Vickrey (1964) and Beckmann (1972), who studied how homogeneous firms equidistantly distributed over a one-dimensional space compete to attract consumers who are evenly distributed over the same space. Each consumer buys one unit of the good, while traveling costs are proportional in distance. Firm i has only two neighbors located at distance Δ on each side. When the travel rate t takes on a high value, firm i is a local monopoly because it is too expensive for consumers located near the midpoint between firms $i-1$ and i to make any purchase. On the contrary, when t is sufficiently low, each firm competes with its two neighbors for the consumers located between them. The market power of a firm is then restrained by the actions of its neighboring firms. In other words, its spatial isolation gives a firm only local monopoly power, for this firm's demand depends on the prices set by the neighboring firms.

Since a reduction in firm i's price significantly affects the demands of its two neighboring firms, they react by reducing their own prices. The reactions of firms $i-1$ and $i+1$ affect in turn the demands of firms $i-2$ and $i+2$ through a ripple effect. This market structure is, by nature, *oligopolistic* in that each firm is concerned directly with only a small number of competitors, in this case, two. Vickrey (1964) and Beckmann (1972) have shown that the market outcome is given by a unique Nash equilibrium in which all firms charge the same price, $p^* = c + t\Delta$, where c is the common marginal production cost. Therefore, spatial separation endows firms with market power. More specifically, as travel rate t, interfirm distance Δ, or both decrease, firms charge a lower price because competition becomes more intense. In the limit, when travel costs vanish, firms price at marginal cost, as in Bertrand. Vickrey's and Beckmann's contributions went unnoticed. It was not until Salop (1979) that scholars in industrial organization started paying attention to spatial competition models.

Yet, how do firms choose where to produce under spatial competition? Hotelling (1929) proposed a solution in a path-breaking paper in which two sellers choose, first, where to set up their stores along Main Street, which is described by a compact segment, and, then, the price at which they supply their customers. Whereas the individual purchase decision is discontinuous—a consumer buys from only one firm—firms' aggregated demands are continuous with respect to prices, except in the case when one firm undercuts the other. Assuming that each consumer is negligible solves the apparent contradiction between discontinuity at the individual level and continuity at the aggregated level. Hotelling considers a rich setting involving both "dwarfs" (consumers) whose behavior is competitive and "giants" (firms) whose behavior is strategic.

Hotelling's conclusion was that the process of spatial competition leads the two firms to locate back-to-back at the market center. If true, this provides us with a rationale for the agglomeration of firms that sell to spatially dispersed customers. Unfortunately, Hotelling's analysis contained a mistake—when firms are sufficiently close, the corresponding subgame does not have a Nash equilibrium in pure strategies—that invalidates his main conclusion. This led d'Aspremont, Gabszewicz, and Thisse (1979) to modify the Hotelling setting by assuming that consumers' shopping costs are quadratic in distance. This assumption captures the idea that the marginal cost of time increases with the length of the trip. In this modified version, d'Aspremont and coauthors show that any price subgame has a unique Nash equilibrium in pure strategies. Plugging these prices into the profit functions, they show that firms now choose to set up at the two extremities of Main Street. In other words, firms selling a homogeneous good choose to be separated because *geographical separation relaxes price competition*.

In the 1980s and 1990s, a great many papers in industrial organization revisited this setup. The main message is simple: the market center is the most attractive place if the duopolists sell goods that are sufficiently differentiated in vertical or horizontal attributes. When travel costs are low, geographical isolation is no longer a protection against competition. Therefore, since competition is relaxed through product differentiation, firms may locate where the market potential is the highest. Furthermore, under a symmetric and bell-shaped population density, Anderson, Goeree, and Ramer (1997) show that firms move toward the city center where the population density is the highest. When the population density becomes more concentrated around the center, equilibrium locations are closer and prices lower. Therefore, *firms face more competition in return for a better access to a larger number of customers*. We will see in section 3 that the home market effect is built on the same trade-off between proximity and competition.

The two-stage game approach struggles to handle a setting with several firms. In addition, when consumers like variety, market areas overlap. This implies that firms no longer have a natural hinterland, but are able to retain customers from very different market segments. De Palma et al. (1985) propose to model consumers' shopping behavior by means of the *multinomial logit*. In this case, consumers need not buy from the cheapest shop nor from the closest one. Instead, their individual demands are smoothly distributed across firms according to a gravity-like rule. Let $\alpha > 0$ be the degree of dispersion of consumers' preferences expressed by the standard deviation of the random term (up to a numerical constant), while Main Street has a unit length by normalization. If consumers' individual choices are independent, then de Palma et al. have shown the following result.

THE GEOGRAPHICAL CLUSTERING OF FIRMS. *If $\alpha \geq 2t$, then at the Nash equilibrium of the simultaneous game with $n \geq 2$, firms colocate at $y^* = 1/2$ and price at $p^* = c + \alpha n/(n-1)$.*

In other words, when the preference for variety is sufficiently strong (α is high enough) and/or travel costs are sufficiently low (t is small enough), firms locate at the market center and price above marginal cost.

But what happens when travel costs are not low enough for $\alpha \geq 2t$ to hold? Osawa, Akamatsu, and Takayama (2017) have recently revisited Hotelling's model by assuming a finite number of equidistant locations along a circle and a continuum of firms selling differentiated varieties. When firms sell their varieties at the same price, firms are evenly distributed across locations when t is large. As t steadily decreases, firms get agglomerated in a decreasing number of commercial districts whose size rises. Eventually, when t becomes sufficiently small, all firms are agglomerated. We will see in the next section that several models in regional economics lead to a similar prediction.

Spatial competition models deliver an important message: *close competitors matter more to a firm than distant competitors*. These models are also appealing because they are well suited for studying several facets of competition in space. Since firms are indivisible and strive to expand their market share by getting closer to consumers, thereby reducing shopping costs, they encapsulate the fundamental trade-off of spatial economics. Unfortunately, spatial competition models rely on fairly specific assumptions and are developed in partial equilibrium settings. When they are generalized or cast in a general equilibrium framework, they very quickly become difficult to deal with. In particular, showing the existence of a Nash equilibrium turns out to be very problematic. Despite these limits, spatial models have proven to be a powerful tool in empirical works devoted to the retail sector (see Aguirregabiria and Suzuki 2016 for a survey).[1]

2.5 *Which Modeling Strategy to Choose?*

If the aim is to avoid the consequences of the spatial impossibility theorem and the pitfalls of spatial competition models, one has to appeal to monopolistic competition or to perfect competition with spatial externalities. Although there is nothing that prevents our using either modeling approach according to the issue at hand, we find it reasonable to assert that the choice of a particular modeling strategy depends on the spatial scale under consideration. Since their extent is often geographically limited, business and consumption externalities are mainly relevant when studying issues arising on a local scale, like in a city. This is why urban economics assumes perfect competition and abides by technological externalities (section 4). The advantage of using externalities is that they are consistent with perfect competition and constant returns, which implies zero equilibrium profits. When the focus is on issues arising on a global scale, spillovers are likely to be absent. In this context, regional economics relies on the combination of internal increasing returns and monopolistic competition to capture the pecuniary externalities generated by the mobility of production factors (see section 3). Under internal increasing returns, the zero-profit condition of monopolistic competition implies that all firms' revenues are paid to production factors.

3. *The Regional Question*

The welfare of people is equally affected by the mobility of commodities and production factors. A shock in transport or trade costs affects firms' and consumers' locational choices. It is, therefore, crucial to have a good understanding of how these agents react—moving to other places or staying put—to

[1] Spatial competition models are also the backbone of the political science theory of party competition in political science, pioneered by Downs (1957) who built on Hotelling (1929).

assess the full impact of trade and transport policies. In this section, we first discuss how the interregional distribution of activities varies with transport costs and with a few other forces that have not received much attention. We build on the home market effect and the core–periphery model. Next, we analyze how the benefits and costs of new interregional transport infrastructures can be assessed once it is recognized that firms and workers are geographically mobile.

3.1 *What Drives Regional Disparities?*

Standard theories predict a market outcome in which production factors receive the same reward regardless of where they operate. When each region is endowed with the same well-behaved production function, capital (for example) responds to market disequilibrium by moving from regions where capital is abundant and receives a lower return toward regions where it is scarce and receives a higher return. If the price of consumption goods were the same everywhere (perhaps because obstacles to trade have been abolished), the marginal productivity of both labor and capital, hence the equilibrium wages and capital returns, would also be the same everywhere due to the equalization of capital–labor ratios.

However, we are far from seeing such a platonic world. To solve this contradiction, Krugman (1980, 1991) takes a radical departure from the standard setting by assuming that the main reason for the uneven development of regions is that firms operate under imperfect competition and increasing returns. This has been accomplished by combining the Dixit and Stiglitz (1977) constant elasticity of substitution (CES) model of monopolistic competition and iceberg transport technology. Consumers are identical and have a preference for variety expressed by a CES utility function. Unlike spatial competition models discussed above, monopolistic competition rules out strategic interactions by considering a large number (i.e., a continuum) of firms producing a differentiated manufactured good or tradable service. Each variety is produced by a single firm and each firm produces a single variety using a fixed and constant marginal requirement of labor. The mass of firms is determined by free entry and each firm has a unique address. If a variety is moved from A to B and if one unit of this variety must arrive at the destination, $\tau > 1$ units must be sent from A to B. The case in which $\tau = 1$ amounts to admitting that transport costs are zero. The level of transport cost is defined by the missing share $\tau - 1$ having melted on the way multiplied by the price of the variety prevailing in A. Hence, a higher value of τ implies higher transport costs. This ingenious modeling trick, proposed by Samuelson (1954) and called the *iceberg transport cost*, allows the integration of positive shipping costs without having to deal explicitly with a transport sector. Nevertheless, we will see in section 3.5 that using the iceberg cost is not an innocuous assumption.

3.1.1 *The Home Market Effect*

Consider an economy formed by two regions, A and B, K units of capital, and L units of labor. Each individual owns one unit of labor and K/L units of capital. Labor is spatially immobile; region A's population share is equal to $\theta > 1/2$. Capital is mobile between regions, and capital owners seek the highest rate of return; the share λ of capital invested in A is endogenous. Labor markets are local and perfect.

In this setup, the interregional distribution of firms is governed by two forces pulling in opposite directions: the agglomeration force is generated by firms' desire for market access, which allows them to better exploit scale economies and to save on transport costs, while the dispersion force is generated by firms' desire to avoid competition in the product and labor markets (Krugman 1980; Fujita, Krugman, and Venables 1999; and Baldwin et al. 2003). This adds competition to the fundamental trade-off of spatial economics (discussed in section 2), and this becomes the *proximity–competition trade-off*. The solution to this trade-off is not straightforward. Indeed, by changing their investment locations, capital owners affect the intensity of competition within each region. This renders the penetration of imported varieties easier or more difficult, which in turn affects the operating profits made in each market. Since operating profits are redistributed to capital owners, their investment decisions, hence their income, also affect the spatial distribution of demand, which influences the location of firms.

This push-and-pull system reaches equilibrium when the capital return is the same in both regions. The upshot is that the larger region hosts a more-than-proportionate share of firms because that region can produce at a lower average cost and supply a bigger pool of consumers. However, the intensity of competition in the larger region keeps some firms in the smaller region. Due to its size advantage, the larger region not surprisingly attracts more firms than the smaller one. What is less expected is that the initial size advantage is magnified, that is, the equilibrium share of firms λ exceeds θ. Since $(\lambda - \theta)K > 0$, capital flows from the region where it is scarce to the region where it is abundant. This result has been coined the *home market effect* (HME).

It is well documented that firms differ in productivity (see Bernard et al. 2007 for a survey). Importantly, the HME may be generalized to heterogeneous firms, which are sorted across spatially separated markets by decreasing productivity. While the low-cost firm can afford more competition in the larger region, high-cost firms seek protection against the devastating effects of competition from efficient firms by locating in the smaller region (Nocke 2006). The spatial selection of firms thus leads to a productivity gap between regions. Furthermore, the HME still holds, which means that some high-cost firms choose to locate in the larger region with the more productive firms (Okubo, Picard, and Thisse 2010). Using US data on the concrete industry, Syverson (2004) observes that inefficient firms barely survive in large competitive markets, but the magnitude of this effect remains controversial (Combes et al. 2012).

Since the aim of the above contributions is to isolate the impact of market size on firms' locations, it is assumed that wages are equal across regions, although relative wages vary with the level of transport costs. As firms congregate in the larger region, competition in the local labor market intensifies, which should lead to a wage hike in A. Since consumers in region A enjoy higher incomes, local demand for the good rises and this should make this region more attractive to firms located in B. However, this wage hike generates a new dispersion force, which lies at the heart of many debates regarding the deindustrialization of developed countries, i.e., their high labor costs. Takahashi, Takatsuka, and Zeng (2013) have shown in a full-fledged general equilibrium model that the equilibrium wage in region A is greater than the equilibrium wage in region B. Furthermore, the HME still holds. In other words, though the wage paid in the larger region is higher, market access remains critical when determining the location of firms.

How does lowering interregional transport costs affect the HME? At first glance, the proximity of large markets should matter less to firms when transport costs are lower. In fact, the opposite holds true: more firms choose to set up in the larger region when it becomes cheaper to ship goods between the two regions. This somewhat paradoxical result can be understood as follows. Lower transport costs make exporting to the smaller market easier, but lower transport costs also reduce the advantages associated with geographical isolation in the smaller region where there is less competition. These two effects push toward more agglomeration, implying that the smaller region becomes deindustrialized to the benefit of the larger one. The HME is thus prone to having unexpected implications for transport policy: by making the haulage of goods cheaper in *both* directions, the construction of new transport infrastructure may lead firms to pull out of the smaller region. This result may come as a surprise to those who forget that highways run both ways.

Unfortunately, the HME cannot be readily extended to multiregional setups because there is no obvious benchmark against which to measure the "more than proportionate" share of firms (Behrens et al. 2009; Matsuyama 2017). The new fundamental ingredient that a multiregional setting brings is that the accessibility to spatially dispersed markets varies across regions. When there are only two regions, the overall impact can be captured through the sole variation in the cost of trading goods between them. By contrast, when there are more than two regions, any global or local change in the transport network is likely to trigger complex effects that vary in nontrivial ways with the structure and shape of the transportation network, as we will see in section 3.5.

Given this caveat, it is no surprise that the empirical evidence regarding the HME is mixed (Davis and Weinstein 1999, 2003; Head and Mayer 2004; Hanson 2005; Costinot et al. 2019). Intuitively, however, it is reasonable to expect the forces highlighted by the HME to be at work in many real-world situations. Yet, how can this be checked? Although it is hard to test the HME because prices are unobservable, a wealth of evidence suggests that market access is correlated to the level of activities. Starting with Redding and Venables (2004), various empirical studies have confirmed the positive correlation between the economic performance of territories and their market potential, defined as the sum of local GDPs or employment discounted by distance. Redding and Sturm (2008) use the separation of East and West Germany between 1949 and 1990 as a natural experiment to study how the loss of market access for cities in West Germany located close to the border affected these cities' growth. They find that, in the period 1949–90, the population growth of West German cities situated close to the border with East Germany was much lower than in cities further from this border. This is clear evidence of a market size effect. After a careful review, Redding (2013) concludes that "there is not only an association but also a causal relationship between market access and the spatial distribution of economic activity." However, care is needed because the market potential used in most empirical studies is a crude approximation of a region's attractiveness (Head and Mayer 2004).

The HME explains why large markets attract firms. However, this effect does not explain *why* some markets are bigger than others. The problem may be tackled from two different perspectives. First, workers migrate from one region to the other, thus leading to some regions being larger than others. Second, the internal fabric of each region determines the circumstances in which a region accommodates the larger number of firms. In what follows, we consider the former approach; the latter is discussed in section 5.

3.1.2 *Can a Core–Periphery Structure Be a Stable Equilibrium Outcome?*

One of the most natural ways to think of agglomeration is to start with a symmetric and stable world and to consider the emergence of agglomeration as the outcome of a *symmetry-breaking* mechanism. The resulting asymmetric distribution involves spikes that can then be interpreted as spatial clusters. It was not until Krugman (1991) that a full-fledged general equilibrium mechanism was proposed. More specifically, Krugman identified a set of conditions for a symmetric distribution of firms and households between two regions to become an unstable equilibrium in a world that otherwise remains symmetric.

The workhorse of the core–periphery (CP) model is again the Dixit–Stiglitz iceberg model. What distinguishes the CP model from the HME is that workers are now spatially mobile. The difference in the consequences of capital and labor mobility is the starting point for Krugman's paper that dwells on pecuniary externalities. When workers move to a new region, they bring with them both their production and consumption capabilities. More specifically, workers produce in the region where they settle, just as capital does, but they also spend their income there, which is not generally the case with capital owners. Hence, the migration of workers, because it sparks a shift in both production and consumption capacities, modifies the relative size of the labor and product markets in the origin and destination regions. These effects have the nature of pecuniary externalities because they are mediated by the market, but migrants do not take them into account when making their decisions because the impact of each migrant is negligible. Such effects are of particular importance in imperfectly competitive markets, since prices fail to reflect the true social value of individual decisions. Hence, studying the full impact of migration requires a general equilibrium framework, which captures not only the interactions between the product and the labor markets, but also the double role played by individuals as both workers and consumers.

Krugman (1991) adds a genuine dispersion force to the baseline model by considering a second sector, called agriculture, which employs a second type of labor. Farmers are spatially immobile and evenly distributed between the two regions. As a consequence, their demand for the manufactured good is rooted in the region where they live. When the agricultural good can be traded at no cost, the equalization of earnings between regions allows farmers to have the same demand functions for the manufactured good. The resulting dispersion of demand prompts manufacturers to choose different locations because they enjoy a proximity advantage in supplying the local farmers.

At first sight, when workers are mobile, the following two effects should lead to the agglomeration of the manufacturing sector in one region. First, when a region is bigger, the HME implies that this region hosts a more than proportionate share of manufacturing firms, which pushes nominal wages up. Second, the presence of more firms means a wider range of local varieties and, for that reason, a lower local price index—a cost-of-living effect. Accordingly, the real wage increases and the bigger region attracts additional workers. The combination of these two effects gives rise to a process of *cumulative causality*, which fosters the agglomeration of firms and workers in one region—the core—while the other one becomes the periphery.

Even though this process seems to generate a "snowball" effect, it is not clear that it always develops according to the foregoing prediction. The above argument ignores several key impacts of migration on the labor market. The increased supply of labor in the region of destination tends to push nominal wages down. In addition, the increase in local demand for tradable goods leads to a higher demand for labor, but the increased competition in the product market tends to reduce firms' profits, hence wages. The final impact on nominal wages is thus hard to predict. Furthermore, when firms are concentrated in the core, sales in the periphery decrease. A priori, the combination of all these effects may spark a "snowball meltdown," which may result in the spatial dispersion of firms and workers.

The timing of events is as follows. First, workers choose their locations. Second, given the interregional population distribution, production takes place. Turning to the specific conditions for agglomeration or dispersion to arise, *the level of transport costs turns out to be the key parameter*. On the one hand, if transport costs are sufficiently high, the interregional shipment of goods is discouraged, which strengthens the dispersion force. The economy then displays a dispersed pattern of production in which firms focus mainly on local markets. On the other hand, if transport costs are sufficiently low, then firms concentrate in the core. In this way, firms can exploit increasing returns by selling more in the growing region without losing much business in the smaller region. The mobility of labor may exacerbate the HME, since the size of local markets changes with labor migration. The CP model, therefore, allows for the possibility of *convergence* or *divergence* between regions. In particular, regions that were once very similar may become very dissimilar. Krugman's work appealed because regional disparities emerge as a *stable equilibrium*, which is the unintentional consequence of decisions made by a large number of economic agents pursuing their own interests.

To sum up, we have the following result.

THE CORE–PERIPHERY STRUCTURE: *Assume that workers are mobile. When transport costs are sufficiently low, the manufacturing, or tradable service, sector is agglomerated in a single region. Otherwise, this sector is evenly dispersed between the two regions.*

The main message of Krugman's contribution is clear: in the presence of increasing returns, lowering transport costs allows some places to build a comparative advantage by making them bigger than others.

When agents are mobile, supply and demand schedules are shifted up and down in complex ways by workers' relocation. It is no surprise, therefore, that coming up with a full analytical solution of the CP model should be impossible. This is what led Krugman to resort to numerical analysis. Subsequent developments have confirmed Krugman's results, but it has taken quite some time to prove them all. The formal stability analysis was developed in Fujita, Krugman, and Venables (1999), but it was not until Robert-Nicoud (2005) that a detailed study of the correspondence of spatial equilibria was provided.

Krugman's paper triggered a huge flow of research, but the quality of this research is unequal. In what follows, we briefly discuss a few shortcomings of the CP model.

(i) The CP model explains *why* agglomeration arises, but does not predict *where* this happens. Indeed, when transport costs are sufficiently low, the manufacturing sector may concentrate in region A or in region B. A natural way to cope with this problem is to endow one region with a comparative advantage. When the number of A-farmers, say, exceeds that of B-farmers, Sidorov and Zhelobodko (2013) show that the interval of τ-values sustaining agglomeration in region A is always wider than that associated with region B. When the number of A-farmers is sufficiently high, regardless of the transport cost level, there exists a unique equilibrium, which is such that all firms set up in region A. This shows how a comparative advantage biases the market in favor of one region, without necessarily excluding the other region as a nest of the manufacturing sector. In a remarkable paper, Davis and Weinstein (2002) document the resilience of urban agglomerations in Japan. After the massive bombing of Japanese cities in World War II, a city typically recovered the population and share of manufacturing that it had before the war. Even Hiroshima and Nagasaki returned to their prewar growth trends after twenty-five years.

Nevertheless, however small a region's size advantage is, the CP model suggests that this region becomes the core when the snowball effect is at work. Instead of comparative advantage, history would act as a selection device among different equilibria. For example, transport technologies used in the nineteenth century led particular sites in the United States to form "portage cities," i.e., sites where a boat or its cargo was carried over land between navigable waterways. Bleakley and Lin (2012a) find that these cities have mostly retained their economic importance despite the disappearance of their original advantage. In sum, *geography and history interact in complex ways to determine the location of activities*.

(ii) The sudden, discontinuous shift from dispersion to agglomeration is the result of assuming that workers are homogeneous and all react in the same way to marginal variations in real wages, very much as consumers react to a marginal price undercutting in the Bertrand duopoly. Once one recognizes that individuals have different attitudes toward the non-monetary attributes associated with migration, the agglomeration process is gradual and sluggish.

In addition, workers' attachment to their region of origin acts as a strong dispersion force. When markets are sufficiently integrated for the real income gap to fall below the utility loss generated by homesickness, the agglomeration process is reversed. In this case, *market integration fosters first divergence and then convergence*; in other words, the economy develops according to the so-called bell-shaped curve of spatial development (Tabuchi and Thisse 2002). This shows that assuming heterogeneous or homogeneous workers may lead to much contrasted results.

(iii) The welfare analysis of the CP model, despite its simplicity, delivers an ambiguous message. Since competition is imperfect, the equilibrium is suboptimal. However, the inefficiency of the market outcome does not tell us anything about the excessive or insufficient concentration of firms and people in big regions. Neither configuration (agglomeration or dispersion) Pareto dominates the other: under the CP structure, workers living on the periphery always prefer dispersion since they have to import all varieties, whereas those living in the core always prefer agglomeration since all varieties are locally produced. In other words, *market integration generates both welfare gains and welfare losses through the geographical redistribution of activities*.

Workers and farmers in the region that attracts the core have a higher welfare than the farmers in the region that becomes the periphery. Therefore, one may use compensation mechanisms to evaluate the social desirability of a move, using market prices and equilibrium wages to compute the compensation to be paid either by those who gain from the move or by those who are hurt by it (Charlot et al. 2006). When transport costs become sufficiently low, the winners can compensate the losers so that they sustain the utility level they enjoyed under dispersion. This is because firms' efficiency gains are high enough to offset the losses incurred by peripheral workers. In this case, regional disparities are the geographical counterpart of greater efficiency. However, when transport costs take on intermediate values, no clear recommendation emerges. This lack of clear-cut results, even in a simple setting such as Krugman's, may explain why so many contrasting views exist in a domain where there is good reason to believe that the underlying tenets are correct.

3.2 *The Evolution of Regional Disparities: Alternative Approaches*

What do the implications of the CP model become when they are studied within alternative settings?

3.2.1 *Input–Output Linkages*

Moving beyond the Krugman model to search for alternative explanations appears warranted in order to understand the emergence of large industrial regions in economies, which are characterized by a low labor spatial mobility. In this respect, a major shortcoming of the CP model is that it overlooks the importance of intermediate goods. Yet, the demand for consumer goods does not account for a high percentage of firms' sales, often overshadowed by the demand for intermediate goods. Therefore, when making location choices, it makes sense for intermediate-goods producers to care about the places where final goods are produced. Similarly, final-goods producers are likely to pay close attention to the location of intermediate-goods suppliers. This is the starting point of Krugman and Venables (1995). Their idea is beautifully simple and suggestive: *the agglomeration of the final sector in a particular region occurs because of the concentration of the intermediate industry in the same region, and vice versa*. Assume that many firms belonging to the final sector are concentrated in one region. The high demand for intermediate goods in this region attracts producers of these intermediate goods. In turn, the intermediate goods are supplied at a lower cost in the core region, which prompts even more final-goods firms to move to the core. Such a cumulative causality process feeds on itself, so the resulting agglomeration can be explained solely by the demand for intermediate goods, without resorting to labor mobility as in Krugman's setting.

Giving intermediate goods a prominent role is a clear departure from the CP model and allows one to focus on other forces at work in modern economies. To this end, note that once workers are immobile, a higher concentration of firms within a region translates into a hike in wages for this region. This gives rise to two opposite forces. On the one hand, final demand in the core region increases because consumers enjoy higher incomes. As in Krugman, final demand is an agglomeration force; however, it is no longer sparked by an increase in population size, but by an increase in income. On the other hand, an increase in the wage level generates a new dispersion force, which lies at the heart of many debates regarding the deindustrialization of developed countries, i.e., their high labor costs. In such a context, firms are led to relocate their activities to the periphery where lower wages more than offset lower demand. Hence, as transport costs fall, *there is first agglomeration and then dispersion of production*. In sum, economic integration should yield *a bell-shaped curve of spatial development*, which describes a rise in regional disparities in the early stages of the development process and a fall in the later stages.

3.2.2 *The Acquisition of Skills*

In the CP model, farmers cannot acquire the skills that would allow them to find jobs in the manufacturing sector. However, a mechanism similar to that in the CP model is at work when immobile individuals may invest in education. To show how this mechanism works, consider a setting in which the manufacturing sector uses both capital and skilled labor; capital is mobile while labor is immobile. If a few firms move away from their region, the unskilled workers residing in the destination region have a higher incentive to be trained. The pool of skilled workers thus grows, which attracts more firms. The higher income that accrues to the newly skilled shifts the demand for the final product upward. Consequently, the region becomes more attractive for the manufacturing firms, which in turn results in even more workers acquiring the skills needed to be hired by the newly established firms. Once more, although all workers stay put, the process of cumulative causality sustains a snowball effect that involves a growing concentration of firms and skills where spatial mobility is replaced by sectoral mobility through training. As in the CP model, the manufacturing sector is dispersed when transportation costs are high and agglomerated when these costs are sufficiently low (Toulemonde 2006). This result points to one of the main reasons for the existence of spatial inequality: *the uneven distribution of human capital*, an issue that we will encounter again in section 4.

3.2.3 *Technological Progress in Manufacturing*

The foregoing models focus exclusively on falling transport costs. There is no doubt that the transport sector has benefited from huge productivity gains over the last two centuries. However, numerous other sectors have also experienced spectacular productivity gains. This situation has led Tabuchi, Thisse, and Zhu (2018) to reformulate the CP model by focusing on technological progress in the manufacturing sector. In addition, these authors recognize that workers are imperfectly mobile. Migration costs act here as the main dispersion force. In the CP model, since prices and nominal wages converge as transport costs fall, the real wage gap, hence the incentive to move, shrinks. What drives Krugman's result is the change in the sign of the real wage differential when transport costs fall below a given threshold. By contrast, in the paper by Tabuchi and coauthors when one region is slightly bigger than the other, technological progress in manufacturing reduces the labor marginal (respectively, fixed) requirement in the two regions, making the larger region more attractive by increasing wages and decreasing the prices of existing varieties (respectively, by increasing wages and the number of varieties) therein. Workers move to the larger region when productivity gains are strong enough to make the utility differential greater than their mobility costs. Therefore, technological progress tends to exacerbate the difference between the two regions and thus raises the incentive to move from the smaller to the larger region. In short, *technological progress in manufacturing favors agglomeration*.

Since innovations often require skilled workers who are more mobile than unskilled workers, Tabuchi and coauthors show that the core is likely to host the most skilled workers. This sheds light on the following important question: is a higher interregional labor mobility the right way in which to lessen regional disparities? Not necessarily. The standard pro-migration argument heavily relies on the assumption that workers are homogeneous. However, people are not homogeneous: migrants are often the top-talent and most entrepreneurial individuals of their region. When the lagging region loses its best workers, those who are left behind will be worse off. In this case, interregional income and welfare gaps are increasingly caused by differences in the geographical distribution of skills and human capital. For example, the spatial distribution of human capital explains about 50 percent of regional disparities in France (Combes, Duranton, and Gobillon 2008). We will return to this issue in sections 4.1 and 5.1.

3.3 *Beyond the Two-Region Setting*

3.3.1 *The Racetrack Economy*

The dimensionality problem mentioned in the study of the HME also occurs in the CP model. Full agglomeration in a two-region setting does not necessarily mean that the manufacturing sector is concentrated in a single region when the economy is multiregional. Akamatsu, Takayama, and Ikeda (2012) and Ikeda, Akamatsu, and Kono (2012) have studied the case of a racetrack economy where workers and firms are initially located in 2^n regions equidistantly distributed around a circle $\mathcal{C}$; farmers are uniformly spread along $\mathcal{C}$. Transport costs between any two regions thus vary with these regions' relative positions on $\mathcal{C}$. Starting from a transport cost value that is large enough for the even distribution of the manufacturing sector among the 2^n regions to be a stable equilibrium, a gradual decrease in transport costs leads to a pattern in which workers and firms are partially agglomerated in 2^{n-1} alternate regions. As these costs steadily decrease, the CP model displays a sequence of bifurcations in which the number of manufacturing regions is reduced by half and the spacing between each pair of neighboring manufacturing regions doubles after each bifurcation until the manufacturing sector is agglomerated into a single region. Therefore, the process of agglomeration described in the CP model remains valid in the case of several regions distributed over a circular space. However, full agglomeration (full dispersion) arises only for very low (very high) values of transport costs. In between these two polar cases, the space-economy displays richer patterns in which the manufacturing sector is concentrated in a relatively small or relatively large number of regions.

Equally important, during the agglomeration process, some regions decline from the beginning of the market integration process. By contrast, some other regions first attract firms and workers over a wide range of values of transport costs before declining. In other words, *a region may first benefit from decreasing transport costs and then lose firms and population later on*. To put it bluntly, the winners and losers of market integration vary with the degree of integration within the economy.

3.3.2 *The Continuum of Locations*

Another solution to the dimensionality problem is the use of an infinite number of locations in a seamless world, which is the representation used by von Thünen, Hotelling, or Beckmann. The real line is the simplest possible homogeneous space on which spatial patterns that display more or less specialized and agglomerated regions may emerge. In a two-region, two-sector setting, regional specialization and spatial concentration do move together. However, working with an infinite space comes at the cost of models that are technically difficult to solve.

To the best of our knowledge, Rossi-Hansberg (2005) is the first attempt made to reconcile a large number of locations, positive transport costs, and agglomeration and dispersion forces within an integrated setup. Locations are ordered along the interval $X = [0,1]$ and a region is a connected subset of X. There are two sectors, the final and the intermediate, each of which produces a homogeneous good; and two primary inputs, land and labor.

Rossi-Hansberg assumes that intermediate and final producers operate under perfect competition, constant returns, and intraindustry externalities that decay exponentially with distance. Final-good firms need the intermediate good and intermediate-good firms receive final goods in return to pay for labor and land. Labor is perfectly mobile across space and between sectors. Whereas firms consume land, consumers do not. The dispersion force thus stems from competition for land between final and intermediate good producers.

Working with a continuum of locations renders market-clearing conditions nontrivial. For example, the law of motion of a good may be described as follows: when there is trade between two locations, an intermediate location $x \in X$ may be viewed as importing the shipment arriving from $[0,x)$, adding exports or subtracting imports at x, and shipping the resulting volume away from x toward $(x,1]$. In equilibrium, the excess supply of this good at $x = 0$ and $x = 1$ must be equal to zero. In such an environment, the price of a good at x depends on how much of the good is traded and lost over the whole pattern of trade, which is determined through the locations of all firms, which in turn depend on the interplay between transport costs and spillover effects.

If transport costs are very high, Rossi-Hansberg (2005) shows that there is no trade. Thus, each location is an autarky hosting both types of firms. However, when transport costs decrease, more regions are involved in trade relations and regional specialization increases. In particular, if shipping the intermediate good gets cheaper, it is less profitable to produce this good close to final producers, while a growing specialization allows firms to benefit from stronger spatial externalities. Furthermore, the final and/or intermediate sector is spatially concentrated in a single region when transport costs go to zero. Loosely speaking, lower transport costs increase the specialization of regions and the geographical dispersion of industries. These results are to be contrasted with those obtained by Krugman and Venables (1995), where declining transport costs first foster agglomeration and specialization and then lead to dispersion. The main reason for this difference in results lies in competition for land between firms, which is a strong dispersion force (see section 4).

3.4 *Where Do We Stand?*

Technological progress in manufacturing and/or transport tends to foster the agglomeration of economic activities. All approaches appeal to cumulative causality, which thus seems to be the main driver of regional disparities. On the other hand, by ignoring the fact that the agglomeration of activities usually materializes in the form of cities, regional economics overlooks the various costs typically generated by the geographical concentration of activities. However, accounting for these costs may have a significant impact on the conclusions drawn from the CP model. More specifically, Helpman (1998) has argued that decreasing freight costs could trigger the dispersion, rather than the agglomeration, of economic activities when the dispersion force stems from a given stock of housing rather than from immobile farmers. In this case, housing competition puts a brake on the agglomeration process, and thus *Krugman's prediction is reversed*. The difference in results is easy to understand. Housing prices rise when workers move to the larger region, which strengthens the dispersion force. Simultaneously, cheaper transport facilitates trade. Combining these two forces shows how dispersion arises when transport costs steadily decrease. Anticipating what we will see in section 4, lowering the transport costs of goods acts against the gathering of consumers within the same city/region because the lower transport costs ease the various congestion costs for the consumer that are associated with the workings of a city (see Rossi-Hansberg 2005 for a different, but related, mechanism). Hence, what is going on *within* and *between* regions is key to understanding the regional question. So far, the most robust conclusion is that lowering transport costs fosters agglomeration as in Krugman, then leads to redispersion as in Helpman (Tabuchi 1998, Puga 1999, Fujita and Thisse 2013).

3.5 *Does Transportation Matter for the Interregional Economy?*

The supply of transport infrastructure is often presented as the panacea to reduce regional disparities. In this section, we take a step-wise approach to this difficult question. After an illustration of the role of transport networks in the location of economic activity, we define transport costs more carefully, as transportation economics brings a few surprises for regional economists. The next step is to identify the main methodological challenges in estimating the effects of transportation infrastructure on economic activities. We will discuss two types of approaches: a structural equation approach derived from spatial general equilibrium models and a reduced-form approach. Both help to determine the relative importance of the many economic mechanisms at work. In the final step, we turn our attention to the political economy question: is there evidence that the money spent on transport infrastructure is invested wisely?

3.5.1 *Transport Networks: What Is Their Impact?*

We consider three different, but equally strong, arguments.

(i) Given a transport network defined by a set of nodes and a set of links that connect some of them, we study a firm's plant that sources inputs and ships outputs to some nodes on the network. When quantities and prices are treated parametrically, the profit-maximizing location problem simplifies to minimizing total transport costs. Transport rates weakly decrease with distance because of the fixed costs of loading and unloading that are minimized at the nodes. In this case, the optimal site is a node in the network (Hurter and Martinich 1989). In other words, the set of all possible locations along roads reduces to the subset of nodes, i.e., a market town, a resource town, a crossroad. By assigning different degrees of centrality to particular nodes of a transportation network, *the new link may favor some nodes at the expense of others*, which may induce the relocation of some firms.

(ii) Consider a simple spatial competition setting with two firms located respectively at $x = 0$ and $x = 1$ that compete in delivered prices instead of mill prices, as in section 2.4. Firms 1 and 2 produce the same good at constant marginal costs such that $c_2 - t < c_1 < c_2 + t$. A simple, Bertrand-like argument shows that, in equilibrium, each firm supplies the market segment in which it has the lower delivery cost. Therefore, the boundary between the two market areas is located at $x_m = (c_2 - c_1 + t)/2t$. When a highway connecting $x = 0$ and $b_1 > x_m$ is built, the transport rate decreases to $\overline{t} < t$ over $[0, b_1]$, but remains equal to t over $(b_1, 1]$. In this case, the more efficient firm supplies a wider range of customers because the highway allows this firm to be more aggressive on $[x_m, b_2]$. Hence, the highway triggers a shift of customers from the inefficient to the efficient firm. Assume now that the highway connects $b_2 < x_m$ and $x = 1$. Firm 2 now invades firm 1's market area, which implies a shift of customers from the efficient toward the inefficient firm. In sum, *the effect of a regional highway varies with the way it connects firms, while the spatial distribution of firms' outputs depends on the industry's cost conditions*.

(iii) Transport infrastructure was built in East and West Africa to allow former colonies to export their mineral resources to developed countries overseas. Bonfatti and Poelhekke (2017) show that coastal countries endowed with mines import relatively more from overseas and relatively less from neighbors than do landlocked countries with mines because the latter needed to be connected to their neighbors to export overseas. This suggests that the transport networks designed during the colonial period still shape the intensity and nature of trade flows in Africa (see also Jedwab and Moradi 2016). In short, *transport networks built long ago may have a lasting effect on the current location of activities*.

With all this in mind, it is hard to believe that transport networks do not matter for the organization of the space-economy. As transport geographers have long argued (see, e.g., Thomas 2002), the spatial distribution of activities depends on the structure of the transport network through the relative values of freight costs along the shortest routes connecting locations.

What are the main transport issues that interest spatial economists? There are at least two. First, regional economic models have the merit of showing that the performance of the transport sector affects the economy in many ways. Hence, what happens in this sector should have an impact on the space-economy. Having said that, it is remarkable that spatial and transportation economics have developed in a rather unconnected way. Second, there seems to be a chicken-and-egg problem that generates endless debates in the media and in academic journals: does the construction of new transport infrastructure foster regional growth, or is this infrastructure built because the corresponding regions are developing fast? The first question is significantly related to what is meant by transport costs and how they are modeled and measured. Since storage and transport are, to a certain extent, substitutes, what matters for firms is the level of *logistics costs*, which account for both types of costs. Freight costs have received the most attention up to now, but we should keep in mind that trade in services accounts for one-third of world exports and even more for intranational trade. Trade in services calls upon very different transport and communication channels: it can be based on electronic delivery, but it also consists of services supplied between branches of a large firm. As for the second question, it strikes us as mainly an empirical issue that has major policy implications, since the construction of a new transport infrastructure is often presented as the remedy to local backwardness.

3.5.2 *Defining Interregional Transport Costs*

The trade and spatial economics literature recognizes the existence of various types of spatial frictions, but often assumes that an iceberg transport cost is sufficient to reflect the impact of these frictions. When the iceberg cost is added to the CES isoelastic demand functions, it is only the level of the demand functions that matters, not its elasticity. The question we investigate here is the relevance of the iceberg cost assumption in modeling the impact of the transport sector on the space-economy.

To the best of our knowledge, the iceberg cost has never been used in transportation economics. The simplest setup in the literature assumes that freight costs for a given value of the load are described by the lower envelope of several affine cost functions where each function represents the freight cost associated with a transport mode. The intercept of these lines takes on its lowest value for trucking and its highest value for air transportation, while their slopes may vary with the size of the vehicles and the frequency of service. This implies endogenous shipping costs. More generally, transportation economists stress the following effects that are not taken into consideration by the iceberg cost. First, *transport modes*: one needs to distinguish among transport modes since they are differentiated by their market structure and technology. Second, *density economies*: transport rates decrease with the size of shipments. Third, *distance economies*: freight rates also decrease with the haulage distance.

Yet, does not taking these effects into account matter to spatial economists? First of all, countries trade with themselves more than with the rest of the world, while most of the trade and economic geography literature assumes that internal transport costs are nil. Shipping goods between two locations within the same country or between two countries involves different costs. Since labor and capital are more mobile within than between countries, the same decrease in transport costs is associated with different responses in firms' and workers' locations. Somewhat surprisingly, the literature devoted to this topic is meager. For example, Behrens et al. (2007) revisited the CP model in which each country is formed by two regions. These authors show that the welfare impact of trade liberalization depends on the internal geography of the two countries. Accounting for density economies renders the two internal geographies interdependent. Indeed, the organization of activities within each country affects the volume of trade, which in turn changes the level of transport costs, and hence the distribution of activities within each country. In particular, the transport policy of a national government may have a marked impact on the internal geography of its trading partners. This makes a case for international cooperation when designing transport policies.

Second, when combined with CES preferences, the iceberg cost assumption implies that halving the producer price of a good entails its delivered price being also cut by 50 percent. In other words, the pass-through is 100 percent. This conflicts with the theory of spatial pricing where freight absorption allows for the penetration of remote markets, as well as with the empirical evidence that suggests an incomplete pass-through (e.g., De Loecker et al. 2016). Thus, iceberg costs cannot capture the richness of the pricing patterns adopted by firms. However, the main results discussed in sections 3.1 and 3.2. hold true in a setting with linear demand functions and additive transport costs, which implies freight absorption (Ottaviano and Thisse 2004). Third, treating freight rates as exogenous amounts to assuming that the transport sector is a perfectly competitive black box; recognizing that freight rates are endogenous affects the location of economic activity. Conversely, trade volumes change with firms' locations, which affects the freight rates set by oligopolistic carriers. In other words, *the location of economic activity depends on carriers' behavior, which itself depends on the way firms and workers are* *distributed across space* (Behrens, Gaigné, and Thisse 2009). Combes and Lafourcade (2005) for France and Winston (2013) for the United States showed that deregulation has been key in the freight rate decrease.

Last, the iceberg cost function cannot account for density economies since the value of the iceberg transport cost τ_{ij} between regions i and j is a constant given a priori. On the other hand, when the number of regions is finite, the iceberg cost function can take distance economies into account, since the values of τ_{ij} are chosen arbitrarily. However, the use of this function in continuous location models, as in Fujita, Krugman, and Venables (1999) and others is more problematic. When one unit of a variety is moved from i to j, only a fraction $\exp(-tD)$ arrives at destination, where $t > 0$ measures the intensity of the distance–decay effect and D the distance between i and j. Therefore, we have $\tau \equiv \exp(tD)$. The unit transport cost is equal to the price of the good at i times the quantity lost en route, which is equal to $\tau - 1 = \exp(tD) - 1$. Hence, the continuous iceberg cost function is increasing and convex in D. This has an odd implication that went unnoticed previously: since the marginal transport cost increases with distance, breaking up a trip into several shorter and connected segments over the distance D is less costly. In other words, using such a function amounts to assuming that there are distance diseconomies. Furthermore, as transport costs tend to increase quickly, remote markets do not matter that much. It is, therefore, not very surprising that the simulation of a shock affecting a region is found to be very localized, e.g. in Hanson (2005).

To sum up, gathering all spatial frictions generated by trade into a single iceberg trade cost is not an innocuous assumption; the main reason is that the level of unit transport costs is endogenous to the spatial structure of the economy. Somewhat ironically, while economic geography stresses the importance of increasing returns in manufacturing, it sets aside the fact that transportation features even stronger economies of scale (Mori 2012). Accounting for the presence of different types of scale economies in the transport sector should rank high on the agenda of spatial economists.

3.5.3 *Does Transport Infrastructure Stimulate Regional Economic Activity?*

Redding and Turner (2015) identify two major methodological challenges. The first is the chicken-and-egg problem, as regions with high transport needs are likely to receive infrastructure. In this case, the construction of a new transport infrastructure is endogenous, rather than exogenous. The second one is the identification of the creation of new activities in the region where the infrastructure is built against the displacement of activities from other regions. Solving the second problem requires special attention for the relocation mechanisms that may be at work. Take an infrastructure measure that targets the trade link between A and B. The increase in economic activity in A can result from a genuine net increase in A, but may also result from a relocation from B to A and/or from a relocation from another region, C, to A. The solution will be to estimate simultaneously the effects of the change in the costs of one transport link on all the regions.

Different approaches are used and their pros and cons are well known (Holmes 2010). There is the *general equilibrium approach,* whose advantage is that it can take into account all the *direct* and *indirect* effects associated with a new transport infrastructure. For example, a location that is not directly affected by new transport infrastructure can be indirectly affected through the redistribution of labor associated with the decrease in transport costs along some least-cost routes. This approach has been developed in the framework of quantitative spatial economics, which we discuss in section 5.2, as it spans both regional and urban economics.

The *structural equation approach* is also embedded in general equilibrium, but only a few equations derived from a general equilibrium setting are estimated. The advantage of this approach, compared with the quantitative spatial approach, is that it is estimated, rather than calibrated. The behavioral parameters are specific to the problem at hand, including their confidence intervals, rather than being borrowed from the literature. The disadvantage is that the structural equation approach is less exhaustive than the general equilibrium one. The third approach is the *reduced-form approach*, which requires less data and resources. However, the causality link is sometimes difficult to assess, so that the main challenge is probably constructing the appropriate counterfactual for the absence of planned transport infrastructure. Moreover, the reduced-form approach rules out the possibility of measuring welfare effects.

In what follows, we discuss the results of the structural equation and reduced-form approaches. Papers are diverse and difficult to compare, as they test the impact of different transport modes (railways, highways, air transport, and high-speed railway) in different periods (nineteenth versus twentieth century) for different sectors (agriculture or manufacturing), as well as in countries having reached different levels of economic development. Since this is an expanding field, we review only a few of the main papers (see Redding and Turner 2015 for a more comprehensive survey).

In two rich, meticulous papers, Donaldson (2018) and Donaldson and Hornbeck (2016) study the effect of the development of railroads in colonial India (1870–1930) and the United States (1870–90), respectively. Railroads in India decreased transport costs and increased agricultural output in the connected districts by 17 percent. Since railroads allowed the different regions to exploit the gains from trade, there was also an overall increase in income for India. In a related way, Donaldson and Hornbeck (2016) aim to quantify the *aggregate* impact of the US railway on the agricultural sector in 1890. The development of the railway system from 1870 to 1890 has increased, directly or indirectly, the accessibility to a growing number of American counties, and has affected accordingly the value of agricultural land rent. More specifically, Donaldson and Hornbeck build on Eaton and Kortum (2002) to determine a market-access reduced-form measure and develop an empirical strategy to estimate how a better access to counties has fostered higher agricultural land rents therein. They find that removing all railroads in 1890 decreases the total value of US agricultural land by 60.2 percent, that is, 3.2 percent of US GDP in 1890, which suggests again a high return for this investment. Unlike manufacturing, agriculture is a dispersed activity that uses a large amount of land. It is, therefore, not terribly surprising that crops located along, or close to, tracks benefited from the construction of railroads.

Berger and Enflo (2017) analyze the effects of 150 years of railways on urban growth in Sweden. They find that the connection to the railway gave cities a strong increase in population in the first twenty years. Later on, the development of railways was much less effective in spurring population growth in the connected cities. Urban growth was then largely due to a relocation effect from the nonconnected towns to the connected towns. As the relative urban growth effect of the initial railroad development shock in the nineteenth century was maintained in the twentieth century, there was again evidence for path dependence regarding the impact of transport networks on spatial development.

Chandra and Thompson (2000) conduct an analysis of the impact of interstate highways on US rural counties between 1969 and 1993. Because the interstate highways were not planned to connect the rural counties, the interstate highway can be seen as an exogenous variable. They find that a new interstate highway increases total earnings in the rural counties the highway goes through. However, it tends to cause a decline in total earnings in adjacent counties. This is mainly evidence for new infrastructures causing a relocation effect of economic activity, rather than a net growth effect. Duranton, Morrow, and Turner (2014) study the impact of intercity trade effects of highways in the United States. They find that highways have a large influence on where production takes place. However, there is no large effect on the value of production. Thus, highways could divert economic activity to depressed areas but would not generate net growth effect.

Storeygard (2016) analyzes the growth effect of the variation of road transport costs between the main port cities and 289 hinterland cities in sub-Saharan Africa. The oil price increase from $25 in 2002 to $97 in 2008 serves as an external shock to transport costs. Storeygard finds an elasticity of −0.28 between the level of road transport costs and urban economic activities whose variations are determined via nighttime light satellite data. However, this elasticity measures only the short-run effects of trade flows.

Studies on China are interesting, as there is both a strong growth in economic activity and an increasing supply of transport infrastructure. However, results for China may differ from other countries because there are still restrictions on labor mobility. Faber (2014) and Baum-Snow et al. (forthcoming) show that the construction of new highways in China increased the industrial output of the connected metropolitan areas and decreased that of the in-between regions. The unconnected regions were less affected. In other words, trade integration reinforces core cities at the expense of intermediate regions, which is consistent with the prevalence of distance economies in the transport sector and of increasing returns in manufacturing sectors.

Reductions in travel costs also affect the spatial equilibrium through the organization of firms. Indeed, firms are packets of functions, such as management, research and development (R&D), finance, and production, which need not be located under the same roof. Before the emergence of new information and communication devices and the development of airlines and high-speed railways, firms that delivered services to other regions relied on local representatives, while headquarters of multiplant firms had local managers to whom they delegated decisions. Lin (2017) finds that Chinese intercity high-speed railway has led to a significant hike in the number of cognitive jobs in connected cities, and Charnoz, Lelarge, and Trevien (2018) use the development of the high-speed railway network in France to show how the decrease in passenger travel time between headquarters and affiliates has allowed management functions to be concentrated in headquarters. Blonigen and Cristea (2015) use the deregulation of the aviation sector in the United States to identify the effect of provision of air services on regional growth. They find that the increase in air services has a significant effect on regional growth, with service and retail experiencing the strongest growth effects.

As noted previously, we find the conclusions that can be drawn from these papers difficult to compare. The papers differ not only by their estimation procedures, but they also focus on different periods, different transport modes operating under different technologies, and different sectors. We need more work accounting explicitly for these differences before providing a clear-cut answer about the net impact of transport infrastructure. Even though it is reasonable to believe that the railway was useful for the development of US and Indian agriculture in the nineteenth century, this does not mean that building new highways reduces today's regional disparities within postindustrial countries. Admittedly, a transport infrastructure, or a few of them, may have a strong impact on the location and growth of activities in the corresponding regions. However, many transport infrastructures are likely to have no impact, as the quasi-ubiquity of these infrastructures will no longer affect firms' locations. In brief, though the provision of more efficient transport infrastructure may help to promote regional growth, we do believe that it does not constitute the universal panacea promoted by many policy makers.

3.5.4 *The Political Economy of Transport Infrastructure*

Transport costs are also the result of political decisions regarding the nature of the transport system. In the United States, the federal government finances a large share of interstate highways using revenues from the gasoline tax. Knight (2004) finds that the workings of the transport committees in Congress allowed building majorities for a regional allocation of funds, which was very inefficient: about half of the investment money was wasted. In addition, the econometric evidence provided by Baum-Snow (2007) and Duranton and Turner (2012) suggests that highways tend to be allocated to cities that grow more slowly than a randomly selected city.

Glaeser and Ponzetto (2018) study the cycle of transport investments in the United States. The United States invested heavily, until the late 1950s, in big urban projects. This was followed by a period of low investment because of social opposition of the "not in my back yard" style in the major urban areas. In the last twenty years, urban investment projects have to demonstrate that all communities suffering the nuisance of the investment are somewhat compensated by large abatement efforts before they can go through. Glaeser and Ponzetto give the example of the Anderson Memorial Bridge that connects Cambridge with Boston that took a year to build in 1915, but over eight years to rebuild a century later mainly because it had to deal with inhabitants living in a well-educated and dense environment. The inefficiency of investment expenditures is the result of political inefficiencies that go beyond the transport committee bias of Knight (2004).

A first inefficiency is that the cost of federal financing for public goods that generate mainly benefits in a city are not very visible, so there is a tendency to overinvest because every city only perceives $1/n$ of the cost, where n is the number of cities. The second inefficiency is due to the opposition power of local groups that suffer the direct nuisance from the investment project. As their damage is direct and the benefits are dissipated, these opposition groups push investment projects to become too costly, as they are forced to include excessive abatement. A third inefficiency is the federal funding rule that is not sufficiently geared to the needs of densely populated cities and cities whose transport infrastructure is used intensively by nonresidents.

In the European Union, federal funds for transport investments are one of the main instruments used by the European Commission for its regional policy. In the late 1990s, the European Union launched a large transport infrastructure program with thirty priority projects. An ex ante assessment of this package yields three main findings (Proost et al. 2014). First, only twelve out of the twenty-two projects pass a simple cost–benefit analysis test. Second, most projects benefit only the region where the investment is to take place, so that the positive spillover argument does not seem to warrant the investment. Finally, the projects do not systematically favor the poorest regions. To conduct the cost–benefit analysis, Proost et al. rely on the spatial general equilibrium model of Bröcker, Korzhenevych, and Schürmann (2010). In this model, the EU has 236 continental regions and each region produces a differentiated and tradable good. Labor is spatially immobile, an assumption that contrasts with that of perfect mobility made in many papers. A node of a transport network represents a region and all regions are connected. The network contains data for all major links in the European road, rail, ferry and air transport networks, including their specific characteristics such as speed limits and likelihood of congestion. Transport cost calculations are based on shortest route through the geographic information system transport network database. New or upgraded links affect trade flows and local production, as well as goods and factor prices. In addition, the model accounts for the funding of investments when assessing households' welfare. We consider this model—which has a rich description of the transport sector and includes several general equilibrium effects—as one of the forerunners of the quantitative spatial modeling approach discussed in section 5.2.

De Rus and Nombela (2007) use a cost–benefit analysis to determine the level of demand needed to make a high-speed railway (HSR) socially beneficial. They find that a link needs some 10 million passengers per year. Many new HSRs in the European Union do not meet this target. When an HSR has to cover all its costs, there will be an insufficient number of passengers for most projects to be economically viable.

The above findings illustrate the role of political economy factors in the selection of projects. As many large projects have dispersed local benefits, some of the misallocation of federal funds can be avoided by relying on local user pricing. Whenever the local user pricing cannot discriminate between local voters and nonvoters while the revenues of user pricing have to be invested in local infrastructure, local user pricing techniques can be more efficient than federal funding (De Borger and Proost 2016). In sum, using federal or European Union funds to finance local transport infrastructure is, at best, a mixed bag.

4. *Why Do Cities Exist?*

The above section addresses the issue of regional imbalance by using ideas and concepts borrowed from the trade and geography literature. In this section, we recognize that cities are often the engine of regional development and turn our attention to the subfield of urban economics where land and spatial externalities are key.

Despite the numerous costs and disadvantages associated with city size, the growing number of urbanites, which characterizes many countries, is evidence that people vote with their feet. Even though urbanization and economic growth do not necessarily go hand in hand, cities are often better places to live than the alternative outlying and rural areas. The main distinctive feature of a city is the very high density of population within a compact area that also accommodates a large amount of buildings and a great variety of infrastructures.

According to Glaeser (2011), the main reason for the existence of cities is to *connect* people. Without the need to be close to one another, how can we explain why in many countries, competition for land gets tougher, as shown by the rising share of housing costs in consumers' expenditures. Households and firms seek spatial proximity because they need to interact for a variety of economic and social reasons. In particular, as new ideas are often a new combination of old ideas, connecting people is crucial for the Schumpeterian process of innovation to unfold. This need has a gravitational nature in that its intensity increases with the number of agents set up nearby and decreases with the distance between them. However, even though people prefer shorter trips to longer trips, they also prefer having more space than less space. Since activities cannot be concentrated on the head of a pin, firms and households compete for land within an area that has a small physical extension, compared with the large regions that are the focus of regional economics.

As shown by Beckmann (1976), individuals' desire to interact with others in a stable and enduring environment may be sufficient to motivate them to cluster within compact areas where they consume relatively small land plots. Beckmann's contribution already tells us how the fundamental trade-off of spatial economics highlights the reason for cities: the population distribution is the outcome of a tension between the propensity to interact with others through a variety of mechanisms that have the nature of benefits external to firms and consumers—the agglomeration forces—and various spatial frictions and crowding effects associated with population size—the dispersion forces. The size of a city is determined as the balance between these forces.

While regional economics focuses on internal increasing returns and the shipment of commodities, urban economics emphasizes the trade-off between increasing returns external to firms and workers' commutes. In the next subsection, we analyze more in depth what we mean by agglomeration economies. Next, we discuss the main urban centrifugal forces, that is, commuting and housing costs. We conclude this section with an analysis of transport congestion and the potential of different urban transport policies.

4.1 *Agglomeration Economies*

Why do consumers choose to live in big cities where they pay high rents, bear long commutes, live in a polluted environment, and face high crime rates? It is due to the much better pay in large cities than in small towns. Yet, why do firms pay their employees higher wages? If firms do not bear lower costs and/or earn higher revenues in large cities, they should rather locate in small towns where both land and labor are much cheaper. The reason for the *urban wage premium* is that the productivity of labor is higher in larger cities than in smaller ones, and labor productivity is higher since a great number of advantages are associated with a high density of activities.[2] These advantages are encompassed under the name "agglomeration economies." They involve both pecuniary and non-pecuniary external effects while they can be intraindustry or interindustry.[3] For a long time, agglomeration economies were used as black boxes hiding rich microeconomic mechanisms that lead to increasing returns at the aggregate level. These boxes have been opened and we now have a much better understanding of these various mechanisms, though their relative importance remains an unsolved empirical question.

In what follows, we distinguish between agglomeration economies that accrue first to firms, and then to consumers.

[2] See Glaeser and Maré (2001), Moretti (2012), and Diamond (2016) for the United States; Combes, Duranton, and Gobillon (2008) for France; Mion and Naticchioni (2009) for Italy; and Gibbons, Overman, and Pelkonen (2014) for the United Kingdom.

[3] The idea of agglomeration economies dates back to Marshall (1890). Intraindustry economies are also called localization economies or Marshall–Arrow–Romer (MAR) externalities; interindustry economies are called urbanization economies or Jacobs externalities. This cornucopia is sometimes a source of confusion.

4.1.1 *The Nature of Business Agglomeration Economies*

Agglomeration economies appear in very different guises. It is, therefore, convenient to organize the various mechanisms associated with population density in the following three categories: *sharing*, *matching*, and *learning* (Duranton and Puga 2004). Their common feature is that they all lead to an aggregate production function displaying increasing returns.

(i) Sharing primarily refers to local public goods that contribute to enhancing firm productivity, such as facilities required by the use of new information and communication technologies and various transportation infrastructures. Sharing also refers to the access to a large pool of specialized workers and to the wide range of business-to-business services available in large cities. Even though firms outsource a growing number of activities to countries where labor is cheap, they also use specialized services available only where these services are produced.

(ii) Matching means that the number of opportunities to better match workers and job requirements, or the suppliers and customers of business-to-business services, is greater in a thick market with many different types of workers and jobs than in a thin one. This idea was formalized almost thirty years ago by Helsley and Strange (1990). Since they face a large number of potential employers, workers living in large cities do not have to change places to switch to another employer. This makes workers more prone to changing jobs. Therefore, workers with the same skills earn higher wages in larger cities, since firms have less monopsony power (Manning 2010). A larger labor market also raises the job seeker's chances of finding employment and makes workers less prone to changing occupations (Di Addario 2011, Bleakley and Lin 2012b). Last, a better match allows workers to focus more on their core tasks in large, rather than in small, cities (Kok 2014).

(iii) Learning in cities may come as a surprise to those who believe that the new information and communication technologies have eliminated the need to meet in person. When different agents possess different bits of information, gathering them generates knowledge spillovers, which is a shorthand expression for the external benefits that accrue to people from the proximity of research centers, knowledge-based firms, and high-skilled workers. As ideas are, by nature, intangible goods, one would expect the internet to play a major role here. Observation shows, however, that research and innovation are among the most geographically concentrated activities in the world (Feldman and Kogler 2010). This is only seemingly a paradox. To be sure, once research has produced new findings, they can be distributed worldwide at no cost. However, the effect of proximity resurfaces when it comes to the creation and acquisition of knowledge (Glaeser 1999, Leamer and Storper 2001).

R&D often demands long periods of exchange and discussion, during which knowledge is gradually structured through repeated trial and error. Thus, extensive, repeated informal contacts between agents located close to one another facilitate the diffusion of new ideas and raise the level of coordination and trust. Although the web is probably the best available library, the profusion of information it offers makes it hard to pick up the few bits of information that are relevant. Learning from other people is often the easiest way to know what is going on and where to search. Hence, innovation would be geographically concentrated because greater creativity is possible when researchers gather. For example, Greenstone, Hornbeck, and Moretti (2010) find that locating a large new plant in a region increases the productivity of other plants in that region. Different works point in the same direction, i.e., *the spatial extent of knowledge spillovers is often limited* (Arzaghi and Henderson 2008, Belenzon and Schankerman 2013, Buzard et al. 2017).[4] Furthermore, spillovers are as local in the 1990s as they were in the 1980s (Lychagin et al. 2016). Thus, contrary to Cairncross (2001), who argued that "new ideas will spread faster" so that "poor countries will have immediate access to information that was once restricted to the industrial world," the empirical evidence concurs with Glaeser (2011) for whom, even in the era of the internet, "ideas cross corridors and streets more easily than continents and seas." There is nothing new under the sun here. Economic historians such as Hohenberg and Lee (1985) have long highlighted the fact that information is one of the main reasons why cities exist. Those results also substantiate the idea that moving people is more costly than moving goods.

Even though the empirical evidence is still thin, knowledge spillovers tend to benefit skilled workers more. In larger and more educated cities, workers exchange more than in cities populated by less-skilled workers. This is confirmed by Rosenthal and Strange (2008) who find that adding 50,000 college-educated workers within 5 miles would increase a college-educated individual's wage by roughly 6 to 12 percent. Bacolod, Blum, and Strange (2009) and Combes, Démurger, and Li (2015) observe that the urban wage premium associated with large cities stems from cognitive skills rather than motor skills. Thus, like in endogenous growth theory, everything seems to work as if the marginal productivity of a skilled worker would increase with the number of skilled workers around him (Moretti 2004, Glaeser and Resseger 2010). However, there would be a rapid geographical attenuation of the positive effects generated by the concentration of human capital.

In a world that is becoming more and more information intensive, the value of knowledge and information is higher than ever for certain economic activities. As a consequence, cities should still be the best locations for information-consuming activities, especially when firms operate in an environment of rapid technological change and fierce competition. Thus, cities specialized in high-tech industries attract high-skilled workers, who in turn help make these places more successful. Therefore, confirming what we saw in section 3, *spatial inequality increasingly reflects the differences in the distribution of skills and human capital across cities*. The flip side of the spatial sorting of workers is the existence of stagnating or declining cities trapped in industries with a limited human-capital base, which are associated with low wages and few local consumer businesses. Thus, even if the spatial concentration of human capital boosts economic and technological development, it might also come with a strong *regional divide* (Moretti 2012). This is likely to have political and social consequences that should not be overlooked: the places left behind might revolt against the ongoing situation through a rise in populist voting (Colantone and Stanig 2018).

[4] International spillovers also decline with distance (Keller 2002). They spread through very different channels, such as international trade and foreign direct investments.

Large cities may be productive for at least three reasons: besides the presence of agglomeration economies such as those discussed above, large cities could host a disproportionate share of highly productive workers and/or of more efficient firms. We briefly discuss the role of workers' sorting below, as well as in section 5.1. As for the selection of efficient firms, Combes et al. (2012) find that there are no significant differences in selection between large and small French cities after controlling for agglomeration economies and workers' sorting.

4.1.2 *How to Measure Business Agglomeration Economies?*

Ever since the seminal work of Sveikauskas (1975) and Ciccone and Hall (1996), research on city size, employment density, and productivity has progressed enormously. Our purpose is not to provide an in-depth discussion of the methodological issues raised by the measurement of agglomeration economies (Rosenthal and Strange 2004; Combes, Duranton, and Gobillon 2011; Combes and Gobillon 2015). Rather, we have chosen to pin down some of the main difficulties that are more directly related to their interpretation.

The basic equation links the average wage w_c in city c to this city's employment or population density den_c:[5]

$$\log w_c = \alpha + \beta \log den_c + \varepsilon_c. \tag{1}$$

An ordinary least squares (OLS) regression yields an elasticity β that varies from 0.03 to 0.09. Hence, doubling the employment density is associated with a productivity increase varying from 2.1 percent to 6.4 percent. However, there are good reasons why these results should be approached with extreme caution since some econometric problems have not been properly addressed.

First, using a simple reduced form such as (1) omits the explanatory variables whose effects could be captured by employment density. For example, overlooking variables that account for differences in, say, average skills or local public goods, is equivalent to assuming that skills or public goods are randomly distributed across cities and are taken into account in the random term. This is highly implausible. One solution is to consider additional explanatory variables, mainly the distribution of skills, the composition of the industrial mix, and the market potential of cities. In doing so, we face the familiar quest of adding an endless string of control variables to the regressions. Using city and industry fixed effects, and individual fixed effects when individual panel data are available, instead allows one to control for the omitted variables that do not vary over time. However, time-varying variables remain omitted.

Second, the correlation of the residuals with explanatory variables, which biases OLS estimates in the case of omitted variables, can also result from endogenous location choices. Indeed, shocks are often localized and thus have an impact on the location of agents, who are attracted by cities benefiting from positive shocks and repelled by those suffering negative shocks. These relocations obviously have an impact on cities' levels of economic activity and, consequently, their employment density. As a consequence, employment density is correlated with the dependent variable and, therefore, the residuals. To put it differently, there is reverse causality: an unobserved shock initially affects wages and thus density through the mobility of workers, not the other way around. This should not come as a surprise; once it is recognized that agents are mobile, *there is a two-way relationship between employment density and wages*. The most widely used solution

[5] Henderson (2003) has pioneered a different approach by checking whether the total factor productivity at the firm level is affected by the density of neighboring plants.

to correct endogeneity biases, whether they result from omitted variables or reverse causality, involves using instrumental variables. This consists in finding variables that are correlated with the explanatory variables but not with the residuals.

Last, the sorting of high-skilled workers into large cities accounts for a large part of the urban wage premium (Combes, Duranton, and Gobillon 2008). Taking into account additional explanations for worker productivity (such as nonobservable individual characteristics or the impact of previous individual locational choices on current productivity) has led to a fairly broad consensus recognizing that, everything else being equal, the elasticity of labor productivity with respect to current employment density is slightly below 0.03. This elasticity measures the static gains generated by a higher employment density (Combes et al. 2012). Recently, De la Roca and Puga (2017) also highlight the existence of significant gains stemming from individuals' accumulation of experience when they work in large cities, which they bring with them when they move to smaller cities. In this case, static estimates also reflect some sort of average of dynamic effects.

It is not disputable that agglomeration economies do exist. However, several issues remain unclear. For example, it is hard to test results that are explained by a specific agglomeration economy, such as those discussed above, and not by another. In a recent comprehensive study, Faggio, Silva, and Strange (2017) give a qualified answer to these questions. They confirm the presence of the various effects discussed above, but stress the fact that *agglomeration economies are the reflection of very heterogeneous phenomena*. For example, low-tech industries benefit from spillovers, though less than high-tech industries. Both intraindustry and interindustry external effects are at work, but they affect industries to a different degree. Firm size also matters: agglomeration effects tend to be stronger when firms are smaller. In other words, specialized and vertically disintegrated firms should benefit more from spatial proximity than larger firms, probably because they have a smaller pool of their own resources to draw from.

Despite the wealth of valuable results, it is fair to say that the dust has not yet settled. If we want to design more effective policies for city development or redevelopment, we need a deeper understanding of the drivers behind the process of agglomeration in cities that vastly differ in size, as well as in historic and geographic attributes. Measuring the relative strength of the various types of agglomeration economies in different urban environments is one of the main challenges that spatial economics faces (Puga 2010, Moretti 2011).

The existence of agglomeration economies has important implications. Once the activities that generate external economies of scale are established, firms and workers tend to become sticky. The cumulative nature of the agglomeration process makes the resulting pattern of activities particularly robust to various types of shocks. This effect is reinforced by the various investments in buildings and infrastructures made by private agents and local governments (Henderson and Venables 2009). So, if there is a priori a great deal of flexibility in the location of activities, agglomeration economies may foster the emergence of a *putty-clay geography*.

Notwithstanding the immense interest of the above contributions, it is worth stressing that urban economists pay little (if any) attention to the market structure of the industry they study, although there are big differences across sectors. For example, the business literature stresses the importance of "co-opetition" where firms share knowledge, but compete fiercely on the product market. Such a market structure fosters a higher productivity and pushes wages upward. Further, when imperfections on the product market (e.g., monopolistic competition) are combined with a perfectly elastic labor supply, workers are underpaid because the marginal productivity of labor is assessed at the firm's marginal revenue rather than the market price.

4.1.3 *Consumer Agglomeration Economies*

Cities are also great places of consumption, culture, and leisure (Glaeser, Kolko, and Saiz 2001). Very much like firms, consumers living in large cities benefit from sharing, matching, and learning through a greater number of tradable and non-tradable goods and services, better transport and communication infrastructures, and a wider array of contacts, cultural amenities, and opportunities for social relations. There is a need to study how the supply of differentiated products affects the welfare of urbanities as a high urban density allows consumers to have access to a wide variety of goods and to save travel costs. A typical example of a non-tradable is provided by restaurants, which account for more than 5 percent of household expenditures. By using tools developed in the study of product differentiation, Couture (2016) shows that consumers are willing to bear substantial travel costs to enjoy their most preferred restaurants, thus confirming the importance of a wide range of varieties as a consumption amenity. A large number of people also facilitate the provision of local public goods that could hardly be obtained in isolation.

The access to a large diversity of goods is a major asset to consumers who have a preference for variety and/or display heterogeneous tastes and incomes. Since a larger city hosts a bigger population of heterogeneous consumers, oligopoly theory holds that markets are more competitive, and provide a broader range of varieties and higher-quality goods (Campbell and Hopenhayn 2005, Berry and Waldfogel 2010, Schiff 2015). As a result, everything else being equal, the market prices of goods should be lower in larger cities than in smaller cities, which is in line with economic geography models where the larger region's price index is lower than the smaller region's. Using barcode data allows Handbury and Weinstein (2015) to compare prices of identical food products available in the same chain store. Their data set includes the prices of hundreds of thousands of goods purchased by 33,000 consumers in 49 US cities. Controlling for product, retailer, and consumer characteristics, they find that prices are almost independent of population size. Handbury and Weinstein also find that the number of available products increases by 20 percent when the city size is doubled. When the price index accounts for this availability effect, groceries are slightly cheaper in larger cities than in smaller ones. For example, when identical products are available in New York City and Des Moines, Iowa, prices in New York are 1.3 percent lower. As for services, the story might well be different. Workers producing consumption services must be paid a higher wage to compensate them for the higher housing and commuting costs they bear in a bigger city. As a result, the price of services should be higher. However, the quality and variety of consumption services is also often higher in larger cities (see Gobillon and Milcent 2013 for a concrete example based on French hospitals).

In short, aside from housing costs, the impact of city size on the cost of living remains an open question. The issue is not purely academic. When housing costs are included, it is reasonable to expect the cost of living to rise with city size. In this case, the fact that federal or national income tax is based on nominal incomes implies a misallocation of workers across cities. Since mobility is driven by differences in real incomes, workers are induced to move from efficient, high-wage cities to less efficient, low-wage cities. Albouy (2009) finds that the resulting misallocation of workers from the North to the South of the United States and away from dense to less dense areas would generate a welfare loss estimated at $28 billion in 2008.

However, consumption amenities may generate perverse effects under the form of inefficient cities, which arise mainly in developing countries. In other words, urbanization need not mean industrialization and trade. Using a sample of 116 developing countries over 1960–2010, Gollin, Jedwab, and Vollrath (2016) find that countries heavily dependent on natural resources are associated with the emergence of "consumption cities," which produce mainly non-tradable services. On the other hand, "production cities," which rely on the development of manufacturing activities and the production of tradable goods, perform much better. In the former case, urbanization is a symptom of the Dutch disease.

4.2 *The Trade-off between Commuting and Housing Costs*

The positive effects associated with city size come with various negative effects such as expensive housing, long commutes, traffic congestion, pollution, and crime. These effects bring about the so-called urban costs. As a result, *cities may be viewed as the outcome of a trade-off between agglomeration economies and urban costs*.

4.2.1 *The Monocentric City Model*

Von Thünen (1826), who studied the spatial distribution of crops around a market town, proposed the first analysis of the way land is allocated across different activities. The authoritative model of urban economics, which builds on von Thünen, is the featureless monocentric city model in which a single, exogenously given central business district (CBD) accommodates all jobs. Locations within a city are thus heterogeneous. Since the only spatial characteristic of a location in this model is its distance from the CBD, the model breaks down the interdependence across location decisions stressed in section 2.2. It is, therefore, compliant with the competitive paradigm. Each consumer uses land in a single location and commutes between her workplace and residence. Therefore, the monocentric city model studies the urban version of the fundamental trade-off of spatial economics, i.e., the trade-off between housing size at one location—approximated by the amount of land used at one location—and the accessibility to the CBD—which is measured inversely by commuting costs.[6]

Consumers, since they dislike long commutes, compete for land with the aim of being as close as possible to the CBD. However, since they also prefer more space, consumers will not cram into the vicinity of the CBD. Therefore, some of them commute over long distances. By allocating to some consumers a plot of land near the CBD, the commuting costs borne by other consumers are indirectly increased as they are forced to set up farther away. Hence, determining where consumers are located in the city is a general equilibrium problem. A perfectly competitive land market may sustain a distribution of identical consumers across locations so as to equalize utility in equilibrium. The argument is disarmingly simple: land users behave as if they were involved in a gigantic auction. Consumers, given their income and preferences, are characterized by a *bid rent function* that specifies the willingness to pay for one unit of land at any distance x from the CBD. A particular land plot is then assigned to the highest bidder. Since the number of consumers is large (formally, a continuum), the winner pays the highest bid and, in consequence, the land rent is the upper envelope of consumers' bid rent functions.

[6] The best synthesis of what has been accomplished with the monocentric city model remains the landmark book by Fujita (1989). Duranton and Puga (2015) is the most detailed and recent survey of urban land use models.

Consider a one-dimensional space $\mathbb{R}_+$ with a dimensionless CBD located at $x = 0$. The opportunity cost of land R_0 is constant and each location is endowed with one unit of land. A mass N of consumers shares the same income Y and the same preferences $U(z,h)$, where z is the quantity of a composite good and h the amount of space used. The price of the composite good, set equal to one, is determined by market forces outside the city. Denoting the land rent prevailing at x by $R(x)$, and the commuting cost borne by a consumer residing at x by $T(x)$, the budget constraint is given by $z(x) + h(x)R(x) = Y - T(x)$.

Let $V(R(x), Y - T(x))$ be the indirect utility. Since consumers are identical, they enjoy the same equilibrium utility level in all locations. As a consequence, the derivative of $V(R(x), Y - T(x))$ with respect to x must be equal to zero. Using Roy's identity, we obtain the Alonso–Muth equilibrium condition:

$$(2) \qquad h(x)\frac{dR}{dx} + \frac{dT}{dx} = 0.$$

Since a longer commute generates a higher cost ($dT/dx > 0$), the land rent must decrease with the distance to the CBD for this condition to hold. As a consequence, (2) means that a marginal increase in commuting costs associated with a longer trip is exactly compensated for by the marginal drop in housing expenditure. To put it bluntly, people trade cheaper land for higher commuting costs. If commuting costs were independent of distance ($dT/dx = 0$), the land rent would be constant and equal to the opportunity cost of land R_0. As a consequence, *commuting costs are the cause and land rents the consequence*. In the featureless monocentric city model, the "something" that explains why the urban land rent exceeds the opportunity cost of land is the physical proximity to the CBD.

Furthermore, the lot size occupied by a consumer must increase with the distance from the CBD. Although a longer commute is associated with a lower net income $Y - T(x)$, the spatial equilibrium condition yields a compensated demand for land that depends on the land rent and the endogenous utility level that is common to all consumers. The utility level is treated as a given by every consumer who is too small to affect it. Since housing is a normal good, a lower price for land therefore implies a higher land consumption. In other words, as the distance to the CBD increases, the lot size increases, whereas the consumption of the composite good decreases. This, in turn, implies that the population density decreases with the distance from the CBD. Hence, lower commuting costs foster urban decentralization.

Despite its extreme simplicity, the monocentric city model tells us something important: when more consumers become agglomerated, *land consumption acts as a dispersion force.* To see how it works, consider the *urban cost* $C(N;u)$ to be borne for the mass N of consumers to enjoy the utility level u. The cost $C(N;u)$ is obtained by summing the commuting costs, the production cost of the composite good, and the opportunity cost of land occupied by urbanites, which are needed for consumers to reach the utility level u:

$$(3) \quad C(N;u) = \int_0^B \left[T(x) + Z(h(x);u) + R_0\right] n(x)\,dx,$$

where $Z(h(x);u)$ is the quantity of the composite good for $U[Z(h(x);u), h(x)] = u$ to hold, while B is the endogenous city limit, and $n(x) = 1/h(x)$ the population density. It can be shown that $C(N;u)$ is strictly increasing and strictly convex in N as well as strictly increasing in u (Fujita and Thisse 2013). Hence, for the utility level to remain the same, the average urban cost borne by the incumbents increases with new residents. In other words, *land use and commuting costs generate agglomeration diseconomies with respect to population size*.

The land rent level reflects not only the proximity to the CBD, but also the "artificial scarcity" of land that stems from restrictive land use regulation, the provision of open spaces, or public policies that maintain the prices of agricultural products far above the international level. For example, the implementation of urban containment hurts new residents by reducing their welfare level or it motivates a fraction of the city population to migrate away (Glaeser, Gyourko, and Saks 2006). In addition, by restricting the population size, such policies prevent the most productive cities from fully exploiting their potential agglomeration effects. Admittedly, environmental and esthetic considerations require the existence of green space. However, the benefits associated with providing such spaces must be measured against the costs they impose on the population. Cheshire, Nathan, and Overman (2014) report that "in 2010, housing land in the South East of England was worth 430 times its value as farmland." We may wonder what shadow price to assign to green spaces to rationalize such a price discrepancy.

After a comprehensive econometric analysis of the various social welfare effects generated by land regulation, Turner, Haughwout, and van der Klaauw (2014) conclude that marginal reductions in land use regulation are likely to have substantial welfare benefits, especially at the edges of existing developed areas. Even more surprising, Hsieh and Moretti (2019) find that lowering constraints on the housing supply in New York, San Francisco, and San Jose to the level of the median American city from 1964 to 2009 would increase the US GDP by 8.9 percent, an astronomical number. It is clear that more work is needed to accurately assess the social cost of the plethora of land and housing regulations. Nevertheless, we may be confident that this cost will be anything but small because land restrictions foster misallocation of labor across cities.[7]

Thus, contrary to a belief shared by the media and the public, the rise in housing costs in many cities is driven mainly by excessive regulation of the housing and land markets. Public policies typically place strong restrictions on the land available for housing and offices. By instituting the artificial rationing of land, these policies reduce the price elasticity of housing supply; they also increase the land rents and inequality that go hand in hand with the growth of population and employment. Since the marginal urban cost $dC(N;u)/dN$ grows, the beneficiaries of these restrictions are the owners of existing plots and buildings. Young people and new inhabitants are the victims of these price increases and crowding-out effects that often make their living conditions difficult, or deter in-migration (Ganong and Shoag 2017).

The monocentric city model has produced a wide variety of results that are consistent with several of the main features of cities. Nevertheless, the basic model of urban economics does not perform so well when it comes to predicting the social structure of cities. For example, when consumers are heterogeneous in income, the model leads to a fairly extreme prediction: households are sorted by increasing income order as the distance to the CBD rises (Hartwick, Schweizer, and Varaiya 1976; Fujita 1989). Hence, the wider the income gap between two households, the greater the distance between their residential locations, and vice versa. This is not what we observe in many countries where metropolitan areas display pronounced U-shaped or W-shaped

[7] A high stock of outdated buildings may also act as a break on city growth. Hornbeck and Keniston (2017) show how the Great Boston Fire of 1872 provided a unique opportunity for new buildings.

spatial income distributions (Rosenthal and Ross 2015). For example, when dwellings are differentiated by age, the affluent are attracted by both suburban locations where they consume big land plots and by downtown locations where central redevelopment makes the housing stock young; this generates a U-shaped spatial income distribution (Brueckner and Rosenthal 2009). In this case, using income-independent commuting costs seems restrictive for studying households' residential choices because there is ample evidence suggesting that the opportunity cost of time increases with income (Small 2012, Koster and Koster 2015). There is a need for more general approaches that account for heterogeneous consumers and differentiated urban environments. What makes such analyses difficult is that the market outcome is given by the solution to a multidimensional matching problem in which housing consumption is income dependent, something that multidimensional matching is so far unable to deal with.

More importantly, the monocentric model remains silent on why jobs would be geographically concentrated. Indeed, firms may alleviate the burden of urban costs in large metropolitan areas through the emergence of secondary employment centers (Henderson and Mitra 1996). To illustrate this, consider the example of a fixed lot size, that is, the consumption of housing is the same across locations ($h = 1$), while commuting costs are linear in distance ($T(x) = tx$). When the city is monocentric at $x = 0$, for a given consumption z of the composite good, the level of urban cost is obtained by integrating $R(x) + tx = tN$ over $[0, N]$, that is,

$$C_1(N, z) = tN^2 + z,$$

while the urban cost becomes

$$C_2(N, z) = \frac{t}{2}N^2 + z$$

when the city has two employment centers located at $x = 0$ and $x = N$, respectively. Since $C_2(N, z) < C_1(N, z)$, the equilibrium utility level is higher in a duocentric city than in a monocentric city. So, we are left with the following question: *Why is there a CBD—or a small number of business districts—in a city?* Scale economies and spillovers are the usual suspects.

4.2.2 *The Emergence of City Centers*

Ogawa and Fujita (1980) tackled this question in a path-breaking paper that went unnoticed for a long time, probably because urban economics was still at the periphery of economics. These authors use a gravity-like reduced form for spillovers and combine consumers and firms in a full-fledged general equilibrium model in which goods, labor, and land markets are perfectly competitive. Spillovers act as an agglomeration force as their intensity is subject to distance-decay effects. However, the clustering of firms increases the average commuting distance for workers, which in turn leads to workers paying a higher land rent. Therefore, firms must pay workers a higher wage as compensation for the higher land rent they have to pay. In other words, the dispersion force stems from the interaction between the land and labor markets. The equilibrium distribution of firms and workers is the balance between these opposing forces. Note the difference with the monocentric city model: interactions among agents make the relative advantage of a given location for an agent dependent on the locations chosen by the other agents, while the agglomeration of firms renders land more expensive at the city center.

Firms produce a homogeneous good using a fixed amount of labor and one unit of land, and a mass N of workers who each consume one unit of land and the final good each. The output level Y of a firm located at $x \in [-b, b]$ depends only on the firm distribution:

$$(4) \quad Y(x) = Y - \int_{-b}^{b} \tau |x - y| m(y) dy,$$

where τ is the distance-decay parameter and $m(y)$ is the density of firms at $y \in [-b, b]$. Ogawa and Fujita (1980) show that the equilibrium urban configuration is unique and monocentric, incompletely integrated or dispersed, depending on the commuting rate t and distance-decay parameter τ. First, when commuting costs are high in relation to the distance-decay parameter, as in preindustrial cities when people moved on foot, the equilibrium involves a complete mix of business and residential activities with everyone living where they work. In this case, land is unspecialized. As commuting costs fall, two employment centers, which are themselves flanked by residential areas, are formed around a district where firms and workers are uniformly mixed. Eventually, when commuting costs are sufficiently low (due to mass transport and the use of cars), the city becomes monocentric. We summarize this as follows.

THE CITY STRUCTURE: *Assume a linear distance-decay spillover and linear commuting costs. Then, there exists a positive constant K such that the city structure is (i) monocentric if $t < \tau K/2$, (ii) incompletely integrated if $\tau K/2 \leq t \leq \tau K$, and (iii) mixed if $\tau K < t$.*

We may rewrite these inequalities in terms of the distance-decay parameter. Hence, the monocentric city emerges when τ exceeds $2t/K$, that is, when the spillovers are very localized. Under (4), O'Hara (1977) shows how the presence of skyscrapers in CBDs may be explained by adding a construction sector to the model. In equilibrium, the building height decreases with the distance from the center of the CBD.

Given the importance of the subject matter, it is surprising that only a handful of papers have explored more general or alternative settings. This paucity makes it hard to adopt a structural approach to studying knowledge spillovers. Fujita and Ogawa (1982) consider a negative exponential decay function and show by using simulations that polycentric configurations exist. One of the main difficulties encountered by these authors lies in the multiplicity of equilibria. Independently, Lucas and Rossi-Hansberg (2002) use a neoclassical production function where land and labor are the two inputs while firms and workers choose their land consumption. Besides a general existence theorem, their simulations yield results consistent with those obtained by Fujita and Ogawa (1982).

Berliant, Peng, and Wang (2002) replace the local firm density by a Lucas-like externality subject to a distance-decay effect. More precisely, a firm produces according to the following Cobb–Douglas production function:

$$Y(x) = A(x) K^{\alpha} L^{\beta},$$

where the total factor productivity $A(x) = [a(x)\bar{K}]^{1-\alpha-\beta}$ is endogenized through the following channels: (i) the aggregate capital stock $\bar{K}$, and (ii) the spatial distribution of firms through the local spillover effect $a(x) = \alpha - x^2 - \beta\sigma^2$ where σ is the absolute deviation of the firm distribution. Berliant and coauthors show that the equilibrium displays one of the three configurations identified by Ogawa and Fujita (1980). The sensitivity of results to the functional form used for the spillovers makes it hard to test their magnitude and is evidence that more work is needed from both the theory and empirical viewpoints.

4.3 *Congestion Costs*

Complaining about transport conditions in cities is as common as talking about the weather. The origin of the discomfort lies in the various negative external costs experienced during most trips. The main external travel costs within cities are due to road congestion, followed by local air pollution, accidents, and climate effects (Parry, Walls, and Harrington 2007). Climate and air pollution are not specific to transport or urban living conditions. These externalities can be tackled efficiently by reducing the emissions per unit of activity: cleaner fuels, catalytic converters, safer cars, and better design of houses. Ignoring crime, the most important negative externality linked to cities is *congestion*. That neither the United States nor the European Union has managed to address these negative externalities efficiently in their pricing policies and infrastructure decisions is probably a major impediment to efficient urban growth. Correcting the external effects generated by urban density and making the best use of agglomeration economies may be one of the main challenges of urban and transportation economics.

People travel within cities for a wide range of reasons, such as commuting to work, business contacts, dropping children off at schools, shopping downtown or in suburban malls, and attending various family and social events. Urban economics focuses primarily on the trade-off between agglomeration economies and the accessibility to the workplace. In accordance with this trade-off, the size and structure of cities are, to a significant extent, driven by the performance of the urban transportation system. To assess the possible impact of various policies, it is important to distinguish between two fundamentally different instruments: a better use of existing transport infrastructure through pricing and regulation, and the addition of transport capacity.

4.3.1 *Reducing Congestion via Road Pricing*

Ever since the pioneering work of Pigou (1932), there has been general (but not universal) agreement among economists that road pricing is the ideal instrument to tackle congestion. The argument is straightforward. Beyond a certain traffic density, travel speed falls as the number of cars increases. Although the costs of travel delays are borne by drivers collectively, each individual driver neglects the external cost of the delays they impose on others. The result is excess travel and inefficiently low speeds. Efficiency can be restored by imposing a toll equal to the marginal external cost.

According to static peak-load pricing theory, this can be explained as follows. Consider two locations, A and B, linked by a road. In the absence of congestion, the cost of a trip is constant and normalized to zero. A population of N homogeneous users residing in A wish to travel to B at the same time, but the capacity of the road is insufficient to allow this. In the simplest formulation, the road has a bottleneck with a flow capacity of s cars per unit time. In this case, the (average) cost of a trip is given by $ATC(N) = \alpha N/s$, where α is the shadow price of travel time. Since the total travel cost is $TTC(N) = \alpha N^2/s$, the marginal social cost of a trip is $MTC(N) = 2\alpha N/s$ and the marginal external cost is $\alpha N/s$. Users internalize the external cost if they pay a toll equal to $\alpha N/s$. Furthermore, if the inverse demand curve for trips is $D(N)$, the optimal number of trips is given by the solution to $D(N) = 2\alpha N/s$. By contrast, the equilibrium number of trips in the absence of a toll is given by the equation $D(N) = \alpha N/s$. The conventional peak-load pricing model thus dictates that, with marginal social cost pricing, the number of trips must be reduced. This creates a potential conflict between reducing traffic congestion and increasing city productivity by maintaining a strong spatial concentration of jobs.

The solution proposed by Pigou is static. However, road congestion is inherently a dynamic phenomenon, as travelers do not have to depart at the same time (Vickrey 1969). Consider the pioneering bottleneck model developed by Arnott, de Palma, and Lindsey (1993). Drivers share the same ideal time to reach their destination and incur a so-called *schedule delay cost* if they arrive earlier or later. Let β denote the unit cost of early arrival (with $\beta < \alpha$) and $\gamma > \alpha$ the unit cost of late arrival. If the rate at which vehicles arrive at the head of the bottleneck exceeds s, a queue develops. The total cost of a trip is the sum of the queuing delay cost, the schedule delay cost, and the toll (if any). Let $\tau(T)$ denote the toll levied at time T, and $N(T)$ the number of vehicles in the queue at time T. A driver who departs at time T incurs a trip cost of $C(T) = \alpha N(T)/s + \beta(\text{time early}) + \gamma(\text{time late}) + \tau(T)$.

Since drivers are homogeneous, the equilibrium travel cost must be the same throughout the period during which drivers depart. If there is no toll, the queuing time increases from zero at the beginning of the travel period to a maximum for the individual who arrives on time, and then decreases back to zero at the end of the travel period. Arnott and coauthors show that the equilibrium travel cost is equal to $\delta N/s$ where $\delta \equiv \beta\gamma/(\beta + \gamma)$. The equilibrium is inefficient because the queuing time is a deadweight loss. Queuing can be prevented by levying a time-varying toll that starts at zero, increases linearly at rate β to a maximum for the individual who arrives on time, and then decreases linearly at rate γ back to zero. In this case, the queuing cost is eliminated, while schedule delay costs are unchanged since the bottleneck still operates at capacity and the interval during which drivers arrive is unchanged. The social cost of travel falls from $\delta N/s$ to $\delta N/2s$, while the private cost inclusive of the toll is the same as with no toll, so that the equilibrium number of trips is still given by the solution to $D(N) = \alpha N/s$. Consequently, with the time-varying toll, the total social variable costs of travel are halved without reducing the number of trips made at all. In other words, the benefits associated with density are not much affected (Arnott 2007). However, congestion costs are not eliminated, since schedule delay costs, which are hidden in Pigou's static model, are unchanged.

So far, we have assumed that car users value time in the same way. However, the empirical evidence suggests that individuals are heterogeneous in their travel time values (Small, Winston, and Yan 2005). In this case, by substituting high-value trips (business and highly skilled commuters) for low-value trips, road pricing generates an additional benefit. By reducing the number of drivers during the peak period, road pricing favors productivity by making business-to-business trips cheaper. Furthermore, individuals are also heterogeneous in terms of schedule delay preferences (Koster and Koster 2015). Combining heterogeneity both in values of time and in schedule delay preferences, van den Berg and Verhoef (2011) show how imposing a time-varying toll at a bottleneck can exploit the heterogeneity of preferences to produce a Pareto improvement even before the toll revenues are redistributed. The issue is important because there is growing evidence that individuals who have a long commute are more prone to being absent from work, arriving late at the workplace, and/or making less effort at work (van Ommeren and Gutiérrez-i-Puigarnau 2011).

Despite these caveats, we may conclude that the smart pricing of a bottleneck can transform queuing into toll revenue; bring about time and productivity gains; be a sensible alternative to the building of new, expensive transportation infrastructures; and dampen the sprawling of activities.

4.3.2 *The Difficult Road to Congestion Pricing*

Congestion pricing has been studied intensively in transportation economics. However, while we have a fairly good understanding of the main issues at stake when agents have fixed locations, the literature does not say much about the locational effects of congestion pricing. Three main lessons can be drawn. First, the design of the road pricing scheme is very important for the magnitude of the total net welfare effects (Anas and Lindsey 2011). For example, Stockholm has implemented a more efficient scheme than has London: the Stockholm system has lower transaction costs and uses more finely differentiated charges depending on the time of the day. Indeed, as shown by the bottleneck model, time differentiation is crucial to capturing the full gains of congestion pricing. A simple differentiation based on day–night, as in London, foregoes a large share of these gains and has to rely mainly on reducing the total number of peak car trips to alleviate congestion. Second, in the 10–20 percent reduction in car use necessary to eliminate most queues, only a relatively small share (40 percent or less) of the suppressed car trips are replaced by mass transit; the rest disappeared due to car sharing, combining trips, or simply foregoing the trip (Eliasson et al. 2009).

Third and last, standard cost–benefit analyses (CBA) are confined to effects within the transport sector, while urban economics suggests that there are wider benefits due to better accessibility that are in line with those associated with higher density. Anderstig et al. (2016) estimate a reduced-form relationship between accessibility and labor income in Stockholm. They find that the total income gains could well be as important as the direct time benefits. Thus, a standard CBA of congestion pricing, focusing only on time benefits, may vastly underestimate the total benefits of substantial improvements in accessibility. Kanemoto (2013) revisits the basic setting of CBA by building a link between spatial and transportation economics. A transport project that reduces commuting costs within a city attracts workers from other cities, so that the additional agglomeration benefit comes at the expense of an agglomeration loss in other cities. In this case, the overall benefits are negative if agglomeration economies in the other cities are larger. By contrast, if the project affects the transport costs of goods produced in imperfectly competitive markets, the additional benefits are always positive and proportional to firms' markups. Thus, care is needed when we come to measuring wider benefits.

In the United States, where road pricing seems to be banned from the public debate, there is more focus on optional varieties of road pricing such as pay lanes. As pay lanes send motorists to an unpriced alternative (the other lanes), they can generate net welfare benefits only when car users differ in their value of time or when there is a fine-tuned pricing of the bottleneck (Small and Yan 2001). Hall (2018) revisits the bottleneck problem when congestion at the bottleneck increases with the queues and shows that a pay lane can lead to a Pareto improvement even before the redistribution of revenues. Indeed, if a sufficiently large number of high-income drivers use the pay lane, congestion is eased on the free lanes. In the European Union, only a few cities (London, Stockholm, Milan, and Göteborg) have implemented congestion pricing schemes. Most national and local governments, alike, favor other policies such as high gasoline prices, large investments in infrastructures, and subsidies in mass transit.

One may wonder why, despite high potential benefits, road pricing is so unpopular. De Borger and Proost (2012) propose a political economy analysis. The population is a priori divided into three categories: nondrivers, drivers who can easily switch to transit (called marginal drivers), and those who face high switching costs. When toll revenues are evenly redistributed across people, nondrivers support congestion pricing. Therefore, nondrivers and marginal drivers could form a majority for congestion pricing. However, if all drivers know only the average switching costs to public transport, marginal drivers expect to bear a cost that might be much higher than what it actually would be. As a result, a majority of the population may be against congestion pricing. However, if road pricing is implemented, the uncertainty is resolved. As a consequence, marginal car users know that their switching costs are lower than what they expected and may therefore support congestion pricing ex post. Hence, a majority of drivers may vote against road pricing ex ante and even against an experiment, since they view their expected gains as being negative, whereas a majority may support it when implemented. As shown by the examples of London and Stockholm, there was an ex ante majority against road pricing but an ex post majority in favor of it. The final decision hinges on the local government's ability to organize such an experiment.

The conventional wisdom holds that, by making commuting more expensive, congestion pricing should lead to denser cities. A handful of recent papers shed a different light on the merits of congestion pricing when it is recognized that agents can change locations in response to substantial changes in travel costs. Using the bottleneck model discussed above, Takayama and Kuwahara (2017) show that densification may not happen when commuters are heterogeneous. On the contrary, the sorting of individuals may lead to equilibria in which the city is less dense. Brinkman (2016) calibrates a spatial general equilibrium model that includes congestion costs and agglomeration economies using data from Columbus, Ohio. Congestion pricing may have negative effects because the congestion tax, although effective in reducing congestion, weakens agglomeration effects through more dispersed employment. Although it seems premature to draw firm conclusions from a small number of papers, it appears that congestion pricing has various unexpected effects that need to be carefully investigated. These papers also confirm one of the main tenets of the survey: a shock to parameters may lead to very different results under fixed or variable locations.

4.3.3 *Does Transport Infrastructure Extension Solve the Congestion Problem?*

Building new road infrastructure raises the value of the road capacity s and reduces the average travel cost $AC(N)$ for any *given* N. However, this argument overlooks the fact that the volume of traffic does not remain the same when road capacity is expanded: the new capacity attracts more car users (N increases). Eventually, expanding the road capacity may create its own demand, a phenomenon known as the Downs paradox (Arnott and Small 1994). This paradox is nothing other than a demand for transportation that is elastic with respect to the level of travel costs. Yet, is this paradox more than an intellectual curiosity? Duranton and Turner (2011) revisited the problem for American cities for the years 1983, 1993, and 2003, while paying special attention to the simultaneity problem between road capacity and traffic density. Their conclusions cast serious doubts on the merits of infrastructure-based congestion policies. First, Duranton and Turner confirm that new roads generate more traffic. More importantly, in the absence of road pricing, they find that "new road capacity is met with a proportional increase in driving." In other words, the elasticity of road use with respect to the number of urban lane kilometers is close to one, thus making the Downs paradox a result. Yet where do the additional travelers come from? Duranton and Turner (2011) find that new cars and new trucks share the responsibility for the extra trips almost equally. In addition, road extension attracts mass transit passengers. This reduces the frequency of public transportation, which in turn increases waiting and schedule delay costs, a vicious circle that may lead to the disappearance of the transit alternative (Arnott and Small 1994). Eventually, roads will attract even more users. This result is not rare in the United States, where car use is cheap and transit alternatives few. In the case of Japan, Hsu and Zhang (2014) find an even slightly higher elasticity of road use with respect to road capacity.

Very much like decreasing transport costs affects the locations of firms, large projects that substantially lower commuting costs are likely to affect household residential choices. For example, Baum-Snow (2007) finds that, between 1950 and 1990, a new highway passing through a central city reduces its population by about 18 percent. His estimates imply that the aggregate central city population would have grown by about 8 percent had the interstate highway system not been built. While central cities were still the origin or destination of 66 percent of all commutes in 1960, this share dropped to 38 percent in 2000. This suggests that jobs have followed residents in their suburbanization. However, care is needed since several effects are at work here (Brueckner 2000). In the same vein, Garcia-López, Holl, and Viladecans-Marsal (2015) studied the effects of highways on urbanization patterns in Spain. They find that, between 1960 and 2011, a highway originating in a central city caused an 8–9 percent decline in the central city population. In addition, a highway fostered a 20 percent population growth in those suburban municipalities where off-ramps were located. Finally, each additional kilometer closer to the nearest highway off-ramp increased municipal density growth by 8 percent. All this confirms the impact of increasing highway capacity on population distribution within metropolitan areas.

These results have two major implications that go against many policy recommendations: when road pricing is not implemented, building new roads may not be the appropriate policy to reduce traffic congestion. In addition, the new roads are likely to have unintended, and possibly undesirable, effects on the urban morphology. Therefore, congestion pricing is back to center stage as the main tool to curb urban congestion. Despite the lack of enthusiasm of policy makers for this instrument, the large number of results obtained by urban transportation economics should encourage governments to assess the merits of smart pricing schemes against the merits of new transportation projects.

Given the long-run implications of transport infrastructure and the externalities they may generate, cities need to be planned. Economists have developed CBA techniques to assess the desirability of transport projects. CBA techniques have progressed over the last fifty years from the Dupuit consumer surplus to methods that correct for externalities and market imperfections, wider economic benefits, and the opportunity cost of public funds. However, CBA methods face great hurdles in assessing these effects correctly (Redding and Turner 2015). Hence, there is a need for broader operational models integrating land use and transport infrastructure (LUTI) with the range of general equilibrium interactions between locations emphasized in theoretical models of economic geography. Unlike most CBA methods, LUTI models are general equilibrium settings where the urban space consists of a finite set of locations. In each location, land can be used for production and/or housing, while locations are linked to each other via transport infrastructure. The LUTI models may also take into account agglomeration economies.

There are two types of models. In the calibrated version, one gathers information on the present equilibrium and complements this with various elasticities taken from the literature to make the model match the present observations. Anas and Liu (2007) provide a good illustration of this approach. Their work may be viewed as a spatial quantitative model of the sort discussed in section 5.2. A good example of the second type of model is by Teulings, Ossokina, and de Groot (2018) who use detailed microdata to estimate a spatial general equilibrium model for the 3000 ZIP codes in the Netherlands. The model allows for changes in population density, job location, and modal choice in an economy with heterogeneous workers and several transport modes. In assessing the impact of a new tunnel in Greater Amsterdam, Teulings and coauthors find that the welfare benefits consist mainly of time savings, but changes in land use may account for up to 30 percent of all benefits.

4.3.4 *What Can Mass Transit Achieve?*

Implementing low prices for urban transit is often presented as a second-best pricing tool that makes up for the missing road congestion pricing. In fact, cheap transit fares have not solved the problem of road congestion; they have created a new one—mass transit congestion. For cheap public transportation prices to help solve the road congestion problem, it must be a good substitute for car use (Parry and Small 2009). In (5), the optimal transit fare is equal to its social marginal cost, corrected by the gap between the price and the social marginal cost of car use. Since the price of an additional car in the peak period is lower than its social marginal cost in the absence of congestion pricing ($P_{car} < SMC_{car}$), subsidizing mass transit is efficient insofar as the subsidy $SMC_{PT} - P_{PT} > 0$ can make car users switch to public transportation. More specifically, for a given subsidy, the fraction φ of new transit passengers who would be car users in the absence of the subsidy must satisfy the following relationship at peak time:

$$(5) \quad P_{PT} \;=\; SMC_{PT} + \varphi \cdot (P_{car} - SMC_{car}).$$

Parry and Small (2009) found that a subsidy close to 90 percent of the average operational costs for urban rail transport is socially desirable when $\varphi = 0.5$. However, empirical studies find values for φ that are often smaller than 0.5. For example, if $\varphi = 0.2$, the optimal subsidy for the peak time drops from 90 percent to 10 percent.

Rail and metro systems display increasing returns because of strong traffic density economies. Hence, first-best pricing gives rise to a huge deficit. This requires a Ramsey–Boiteux pricing scheme that takes into account the opportunity cost of public funds and adds an extra margin for the less-elastic users to further reduce the deficit characterizing almost all public transport systems. Bus systems exhibit smaller scale economies, so that accurate peak load pricing can allow bus systems to break even more easily. Moreover, buses also contribute to road congestion, unless they have a separate lane. Hence, there is a need for more efficient bus pricing systems. They should account for the differences in cost between peak and off-peak trips and vary with the area and distance traveled, as well as with the congestion level of roads. This would increase the overall efficiency of the urban transport system and alleviate the financial problems of urban public transport agencies.

The mix of road and mass transit is the result of a political choice made in a city with heterogeneous users. Brueckner and Selod (2006) show that a voting equilibrium need not yield the best mix of capacities.

5. *Toward a Synthesis of Regional and Urban Economics*

It should now be clear that we need a better integration of different types of spatial frictions discussed in the two sections above to understand how forces acting on different spatial scales shape the economy. Most countries trade more with themselves than with the rest of the world, while large cities are major actors in the process of trade. It is, therefore, fundamental that one understands how the intensity of interregional and international trade is influenced by the size and structure of cities and, conversely, how trade and market integration affect the structure of cities. This task must be accomplished within multi-region frameworks that capture as many general equilibrium effects as possible. In what follows, we consider two different approaches. First, we discuss the formation of urban systems, a natural framework to consider as the performance of regional economies is often determined by the economies of the cities they include. The second approach is the quantitative spatial modeling strategy. At first sight, the two approaches appear to have no elements in common; they are also analytically fairly different. However, each may be viewed as an attempt to take on board different ingredients of regional and urban economics. By mixing the "old" and the "new" in various combinations, they offer new and alternative lines of research.

5.1 *The Urban System, or Why Not All Cities Are Alike*

The most enduring problem in spatial economics is probably the existence of an urban system involving large and medium-sized cities, towns, and villages that produce and trade different commodities and host different types of workers. Two generations of models shed light on different aspects of this problem. The first one focuses on the size and specialization of cities, while the second addresses their skill composition.

5.1.1 *Cities Differ in Size and Specialization*

Despite valuable early contributions by Christaller (1933) and Lösch (1940), it is fair to say that Henderson (1974, 1988) was the first to develop a compelling and original approach that allows the formation of an urban system to be described. His setting involves an endogenous number of specialized cities that trade commodities. Hence, the urban system resembles a wide array of small open economies where firms operate under perfect competition and external economies of scale. The market for cities is characterized by competition among land developers. The land developers understand that they may benefit from organizing cities in a way that maximizes the local land rent, while internalizing the external effects generated by the agglomeration of firms belonging to a particular industrial sector. In equilibrium, the utility level is the same across cities and each city has a positive, finite size. As cities vary in their industrial specialization, they have different sizes, since industries differ in the external economies they can create.

To highlight the forces at work in Henderson's urban system, consider an economy with $n > 1$ sectors, each producing a homogeneous and tradable good by using labor. There is a large number L of identical consumers, each endowed with one unit of labor and Cobb–Douglas preferences $U = x_1^{\alpha_1} \cdots x_n^{\alpha_n} h$, where $h = 1$ stands for the fixed lot size and $0 < \alpha_i < 0$ and $\sum \alpha_i = 1$. Trading goods between cities is costless, so each good is available at the same price across cities. In each city, there is again a tension between external economies associated with the agglomeration of firms at the CBD and the diseconomies generated by the need to consume land and commute to the CBD. The production function of sector $i = 1, \ldots, n$ is given by $F_i(N) = N^{1+\gamma_i}$ where $0 < \gamma_i < 1$ is the degree of increasing returns in sector $i = 1, \ldots, n$; good n is chosen as the numéraire. Goods are indexed for $\gamma_1 > \cdots > \gamma_n > 0$ to hold. Commuting requires tx units of the numéraire to cover distance x. Wages, the land rent, and the number and size of type-i cities adjust to equalize utility across cities.

Fujita and Thisse (2013) show that the equilibrium involves

$$m_i^* = \alpha_i \frac{(1-\gamma_i)^2}{\gamma_i} \frac{t}{4k} L, \qquad i = 1, \ldots, n-1, \tag{6}$$

type-i cities whose size is given by

$$L_i^* = \frac{\gamma_i}{1-\gamma_i} \frac{4k}{t}, \quad i = 1, \ldots, n, \tag{7}$$

where $k \equiv \left(4\gamma_1/t\right)^{\gamma_1/(1-\gamma_1)}\left(1-\gamma_1\right)$ decreases with t. Since urban costs are higher in larger cities than in smaller, wages are higher in the former than in the latter. This reflects the fact that larger cities have a higher degree of increasing returns.

Hence, the equilibrium number of type-i cities decreases with stronger scale economies, but increases with lower commuting costs. Equilibrium city sizes increase with the intensity of increasing returns, but decrease when commuting becomes more expensive. A higher expenditure share on good i leads to a larger number of type-i cities whose size remains the same. Based on (6) and (7), the fundamental trade-off between increasing returns and commuting costs shapes the urban system as follows: (i) when increasing returns become stronger in the production of commodity i, there are fewer type-i cities but they become larger; and (ii) when commuting costs rise, all cities become smaller, but the number of cities of each type increases. Increasing returns prevent the proliferation of cities ($m_i^* \to \infty$ when $\gamma_i \to 0$), while commuting costs prevent cities from being indefinitely large ($L_i^* \to \infty$ when $t \to 0$).

It follows from (7) that large (small) cities are *specialized* in the production of goods with high (low) degrees of increasing returns: $L_1^* > \cdots > L_n^*$. In larger cities, workers are paid higher wages, but the wage difference is offset by higher land rent and commuting costs. Surprisingly, there may be fewer large cities than small because m_i^* depends on α_i. Hence, we need to impose restrictions on the consumption and production parameters to get a pyramidal urban system.

Duranton and Puga (2005) go one step further by recognizing that firms are packages of different functions, such as headquarters and production plants. Due to the development of new information and communication devices, firms are now able to distribute these functions among geographically separated units in order to benefit from the attributes specific to different locations. However, caution is needed. Since production requires both embodied and disembodied knowledge, the transmission of bits of information via e-devices remains incomplete and imperfect (Leamer and Storper 2001). When communication costs between spatially separated headquarters and plants are high, firms remain integrated and the urban system displays the pattern described above. By contrast, when communication costs are sufficiently low, Duranton and Puga (2005) show that headquarters get concentrated in a few large cities where they use a wide array of business-to-business services, while plants are located in specialized and smaller cities. In other words, *cities shift* *from sectoral specialization to functional specialization*.

Fujita, Krugman, and Mori (1999) develop a very different approach that relies on economic geography. There are several tradable goods produced under increasing returns and monopolistic competition. Land is used by farmers to produce the agricultural good, but neither firms nor manufacturing workers use land. Under CES preferences, each good is characterized by a specific elasticity of substitution. Intercity trade is costly, but the unit transport cost is the same across goods. As the total population grows, Fujita, Krugman, and Mori (1999) show that a hierarchical urban system emerges: higher-rank cities provide a larger number of manufactured goods. Hence, *cities are diversified*. In addition, there is two-way trade across cities because cities produce differentiated goods. Horizontal relations among cities are thus superimposed on the pyramidal structure of the urban system. There is also trade between cities and the rural areas where farmers live and work. However, this approach remains in the tradition of the CP model as cities have no spatial extension.

Last, Desmet and Rossi-Hansberg (2013) construct yet another model to understand the formation of a system of cities without appealing to specialization and trade. The population consists of identical and perfectly mobile individuals. The city size distribution is the outcome of the interplay between three effects, i.e., productive efficiency, consumption amenities, and spatial frictions. Amenities are exogenous shocks that affect preferences, while the main friction stems from commuting in a monocentric city. Commuting requires a public infrastructure financed by a tax on labor. The model is calibrated to the US cities with more than 50,000 inhabitants. This model is then used to assess the effects of the aforementioned three components on the size distribution of cities and the common utility level. Eliminating differences in efficiency, amenities, or frictions leads to a significant reallocation of people across cities, but only to small gains in welfare (at most 2 percent). The perfect mobility of labor is likely to play an important role in "re-optimizing" the city distribution and limiting losses in welfare. The same exercise is undertaken for 212 Chinese cities. This also generates significant changes in city size, but the gains in welfare are now substantial, exceeding 40 percent in some cases. One possible explanation for this difference in results might be the presence of substantial constraints on internal migration in China, which prevent the large Chinese cities from reaching their efficient size (Au and Henderson 2006, Bosker et al. 2012). This modeling approach helps one understand how some of the main driving forces of spatial economics interact to determine the city distribution. However, the model is too stylized to trace back the origin of differences in productivity and amenities.

The Zipf law.—The actual number and size of cities seem to obey a simple, but intriguing, empirical rule, which keeps drawing attention. In 1913, the German geographer Felix Auerbach found an unexpected empirical regularity: the product of the population size of a city and its rank in the distribution appears to be roughly constant for a given country. To put it differently, if there is a single type-1 city, the two type-2 cities have half the population P_1 of the largest city, the three type-3 cities host a third of that population, and so on. Denote by P_i the population of cities of rank i in the urban hierarchy. Formally, the Zipf law rule holds that

$$\log i = \log P_1 - b \log P_i$$

with $b = 1$. A high number of estimations of b suggest a value close to 1. Even if a power function provides a good approximation of the population distribution, is $b = 1$ the best estimation? Using an algorithm to determine, in a consistent way, city boundaries with high-resolution geographical data, Rozenfeld et al. (2011) find that the Zipf law is a good approximation of the population distribution for Great Britain and the United States. These authors also find that the same rule applies to the area distribution. The pooled estimate of the coefficient b obtained by Nitsch (2005) in a meta-analysis combining 515 estimates from 29 studies suggests a value close to 1.1! Can such a remarkable result be micro-founded by an urban economics model?

Using the above solution of Henderson's model sheds light on the Zipf law. Multiplying m_i^* and N_i^* yields the total population $P_i = \alpha_i(1-\gamma_i)L$ living in type-i cities. Renumbering the sectors for $P_1 > \cdots > P_{n-1} > P_n$ to hold, we obtain

$$\ln i = \ln\alpha_1 + \ln(1-\gamma_1) - \ln\alpha_i - \ln(1-\gamma_i).$$

Since this expression involves demand- and supply-side parameters that are a priori independent, there is a priori no reason to expect the right-hand side to be equal to the left-hand side.

A considerable effort has been made to explain the Zipf law (see, e.g., Gabaix 1999; Eeckhout 2004; Hsu 2012; Behrens, Duranton, and Robert-Nicoud 2014). The plethora of existing setups relies on heavy, and sometimes ad hoc, models that provide, at best, approximations of the Zipf law. This is the reason why we believe that this law is a mystery and likely to remain so.

5.1.2 *Cities with Heterogeneous Workers*

In Henderson-like models, cities are populated with identical and homogeneous workers. Yet, it is well documented that cities host heterogeneous workers, and thus differ in their respective social composition. In particular, the wage distribution in large cities first-order stochastically dominates that in small cities. So, we need to determine what large and small cities look like when workers are heterogeneous in skills. This is precisely what the new generation of models aims to explain by studying the sorting of heterogeneous workers between and within cities. Importantly, these models explicitly address the issue of human capital whose empirical relevance was stressed in section 4.1. In equilibrium, individuals do not want to change place and occupation, while all markets clear. Unlike Henderson (1974, 1988), these models do not rely on the presence of "large agents" such as developers. In other words, *cities emerge through the interplay between choices made by a large number of small agents*.

Agglomeration, sorting, and selection.— The economy consists of a continuum of monocentric cities and a continuum of individuals who differ in skill $s \in \mathbb{R}_+$. An individual's productivity hinges on two factors: his skill and his matching with a specific city. An individual knows his skill, but the quality c of the match with the city is a priori unobservable. When an individual locates in a city, he observes his own productivity, given by $\varphi = c \times s \in \mathbb{R}_+$, where $c \in \mathbb{R}_+$ is the realization of a random variable. This variable accounts for the environment that determines the quality of the match between the individual and the city. Knowing his productivity, the individual chooses to be an entrepreneur or a worker. Individuals are risk neutral and consume a final good and a fixed lot size. Hence, individuals maximize their consumption of the final good. This good is produced by using a CES bundle of differentiated inputs, where $\sigma > 1$ is the elasticity of substitution across inputs. An entrepreneur sets up a firm that uses labor to produce a variety. Intermediate firms are heterogeneous, compete under monopolistic competition, and their number is equal to the number of entrepreneurs. Under commuting costs linear in distance to the CBD (tx), per capita urban costs are equal to $tL/4$ where L is the city size.

The great merit of the model developed by Behrens and coauthors is that it allows one to (i) separate agglomeration economies associated with city size, the sorting of workers, and the selection of efficient firms; and (ii) study how these forces interact to shape the urban system. It is natural to begin with a city c whose population L and productivity distribution $F(\varphi)$ are given.

Since individuals are heterogeneous, there is a cutoff $\varphi_{\min}$ such that an individual chooses to be an entrepreneur if $\varphi > \varphi_{\min}$, while those with a productivity smaller than $\varphi_{\min}$ become workers. Since the production function is CES, the share of entrepreneurs is independent of the population size L. This provides a rationale for the absence of selection found by Combes et al. (2012) in French cities. Behrens and coauthors show that the production function of the final sector may be rewritten in terms of workers' productivity:

$$(8)\qquad Y = \left[\int_{\varphi_{\min}}^{\infty} \varphi^{\sigma-1} dF(\varphi)\right]^{1/(\sigma-1)} \times \left[\int_{0}^{\varphi_{\min}} \varphi \, dF(\varphi)\right] \cdot L^{\sigma/(\sigma-1)}.$$

The total output per capita is thus proportional to $L^{1/(\sigma-1)}$, as in the case of homogeneous individuals (Fujita and Thisse 2013, chap. 4). Raising L leads to a higher number of intermediate firms, which makes the final sector more efficient and leads firms to pay a higher wage. Hence, there are agglomeration economies associated with population size. The city has a finite size if urban costs rise faster than per capita income, that is, $\sigma > 2$. Assume now that the productivity of each individual of the same population is given by $\lambda\varphi$ with $\lambda > 1$, rather than φ. It follows from (8) that Y/L increases by a factor equal to $\lambda^2 > 1$. However, the cutoff changes with the skill distribution, which is now $F(\lambda\varphi)$. It can be shown that the new cutoff is equal to $\lambda\varphi_{\min}$. Hence, the intermediate sector involves more productive firms, which pay a higher wage, so the entrepreneurs also earn more. In other words, a higher-skilled population generates a human capital externality in which the productivity of an individual increases with that of his coresidents. Behrens and coauthors show how the interplay between these externalities interact for skill and population size to be *complements*, that is, the earnings of workers and entrepreneurs increase with both their individual skill and the city size. This is consistent with De la Roca (2017), who observes that Spanish migrants who move to big cities are positively selected by their educational level and individual productivity.

Since highly skilled individuals benefit from living and working in large and expensive cities while the number of potential cities is large, there exists a unique equilibrium in which cities are homogeneous in skill. However, due to the variations in the quality c of the match, cities host people who differ in productivity, hence workers and entrepreneurs are heterogeneous. The equilibrium urban system displays the following major characteristics. The size of a city increases with the intensity of agglomeration economies ($1/(\sigma - 1)$) and the skill level of its residents (s), but decreases with urban costs (t). Hence, the fundamental trade-off of spatial economics is augmented by a skill effect that magnifies the agglomeration economies. Since two cities sharing the same skill have the same size, it is possible to retrieve their number from the skill distribution. When the cumulative distribution of skills is concave, there are fewer bigger cities. In this case, *the urban system displays a pyramidal structure with the most* *efficient cities at the top and the least efficient at the bottom.*

There is no question that the work of Behrens and coauthors provides a rich description of the urban system as cities differ in skills and accommodate individuals who differ in productivity. What makes this paper unique is that the model identifies conditions implying that skill and population size are complements. That the aggregate productivity of bigger cities exceeds that of smaller cities while the former pay a higher nominal wage than the latter concurs with Henderson (1974; 1988), which is reassuring. Both settings thus provide a rationale for the existence of the urban wage premium. However, unlike Henderson, cities produce the same final good since the production function is the same across cities. Therefore, the urban system is formed by autarkies.

City size and learning.—To the best of our knowledge, Davis and Dingel (2019) are the first who have shown how costly and freely chosen face-to-face contacts across heterogeneous individuals may drive urbanization and the emergence of an urban system. Again, the economy consists of a continuum of monocentric cities and a continuum of individuals who differ in skill. Individuals consume three goods: non-tradable services, tradable goods, and housing. Individuals choose to produce non-tradable services or tradable goods. Urban costs are equal to $tL/4$ where L is the city size. The indirect utility of an individual with income y in city c is $V(s,c) = y - np_c - tL/4$, where $p_c\,(n)$ is the price (consumption) of non-tradables.

Non-tradables are produced under constant returns and the corresponding individual income is city-specific and normalized to p_c. By contrast, the output of an individual producing tradables depends on his skill and his learning opportunities, which are determined by the composition of the local population. Individuals producing tradables can freely divide their available time between learning and producing. They may exchange ideas with and learn from anyone. However, an individual learns more from interacting with more-skilled individuals. Learning, one of the main agglomeration economies discussed in section 4.1, is here endogenous, costly, and city-specific. By increasing the productivity of individuals involved in the exchange of ideas and knowledge, the externality generated by learning plays the role of external increasing returns.

Key to Davis and Dingel (2019) is the idea that individual skills and the quality of the learning environments are complements. More specifically, the concentration of skilled workers magnifies each worker's own productivity advantage. This is the agglomeration force. Obviously, urban costs are higher in bigger cities. Therefore, services are also more expensive because their producers must be compensated for these higher costs, which makes urban costs even stronger as a dispersion force. Producers of tradables will choose to live in big cities where urban costs are high and services expensive if their earnings are sufficiently high. In equilibrium, individuals choose an occupation, a city, and an allocation of time to maximize utility, and all markets clear. Davis and Dingel show that, when urban costs are not too high, there is spatial sorting along the skill dimension: *large cities are more skill-abundant and skill premiums are higher in bigger cities*, in line with the empirical evidence provided by the authors. Otherwise, as in the CP model, stable equilibria have equal-sized cities when the agglomeration force is weak relative to urban costs.

The complementarity between skills.—Eeckhout, Pinheiro, and Schmidheiny (2014) observe that the standard deviation of the skill distribution rises with city size. More specifically, these authors find that the distribution of skills in US cities has thick tails in large cities, i.e., these cities host disproportionately both high-skilled workers and low-skilled service providers.

Hence, the social structure of big cities is more heterogeneous than that of small cities. Eeckhout and coauthors argue that the complementarity between heterogeneous workers drives the spatial sorting across cities. Therefore, the key issue is to determine which combination of skills mutually boosts their productivity: Do the high-skilled workers benefit from the presence of medium-skilled or low-skilled coworkers?

Differences in cities' total factor productivity lead to differences in demand for skills. Using CES-aggregators for the production function with three types of labor ranked by increasing order of skill, Eeckhout and coauthors characterize the market outcome for each of four technology patterns: extreme skill and top skill, each with complements or substitutes. They show that the skill distribution has thicker tails in the larger city under the extreme-skill complementarity (1 and 3). Under the top-skill complementarity (2 and 3), only the highly skilled are disproportionately represented in the larger city. Top-skill and extreme-skill substitutability yields the mirror-image results. All of this pleads in favor of complementarity between high- and low-skilled workers.

The above papers complement each other well. Behrens and coauthors show that a population size externality and a composition externality make skill and population complements, which generate spatial sorting. Davis and Dingel (2019) show that a learning externality magnifying individual productivity gives rise to spatial sorting through big cities displaying a disproportionate and increasing share of skills. According to Eeckhout and coauthors, the spatial sorting of skills is driven mainly by the complementarity between high-skilled and low-skilled workers, which entails big cities hosting disproportionately both high- and low-skilled workers. It is reasonable to expect the three ideas to be part of the story. Combining them within the same setting is another issue.

Unfortunately, apart from Fujita, Krugman, and Mori (1999), the new urban system models still assume that intercity trade is costless or cities produce the same good. In other words, all the settings fail to recognize that cities are anchored in specific locations and embedded within intricate networks of trade relations that are critical in explaining cities' sizes, industrial mixes, and social structures. Therefore, cities are like "floating islands." A comprehensive theory of urban systems should explicitly account for city location. For this, we must assume positive transport costs. This proves to be a difficult task because commodities are sold at different prices in different cities. Furthermore, local and global forces interact to determine the geography of production and employment, as the size of cities depends on the interplay between the emergence of employment centers *within* cities and trade flows *between* cities. This is a task that quantitative spatial modeling aims to accomplish.

5.2 *Quantitative Spatial Economics*

Most economic geography models, by focusing on two symmetric regions, are highly stylized settings. Likewise, by assuming a featureless space with a given CBD, the monocentric city model abstracts from spatial characteristics that affect both consumers' locational choices and the layout of cities. The difficulty in accounting for spatial inhomogeneities and the lack of an analytical solution to the dimensionality problem has led to the development of a new strand of literature, called *quantitative spatial models*.

5.2.1 *What Are Quantitative Spatial Models?*

The aim of quantitative spatial models (QSMs) is to overcome the above limits by developing general equilibrium settings that consider theory seriously (see Desmet and Rossi-Hansberg 2013, Allen and Arkolakis

2014, Ahlfeldt et al. 2015, and Redding 2016 for the pioneering contributions). QSMs take us further than the regional general equilibrium model à la Bröcker, Korzhenevych, and Schürmann (2010) discussed in section 3.5, and even further than the land use models à la Anas and Liu (2007) discussed in section 4.3. More specifically, QSMs take on board the main agglomeration and dispersion forces uncovered in regional and urban economics, such as market access, the relationship between productivity and density, and urban costs. Inevitably, they must leave out other effects, so that the specification of the theory model is a key issue. QSMs also account for a large number of locations that are differentiated exogenously by given amenities and/or productivity differences. Therefore, these models may be viewed as settings that provide rich syntheses of regional and urban economics.

The main purpose of quantitative spatial models is not (necessarily) to provide new theoretical results but to assess public policy interventions or the impact of shocks. To achieve their goal, these models must admit a small number of exogenously given parameters (e.g., the elasticity of substitution under CES preferences), while the other parameters are calibrated on the observed variables, i.e., the data. The model is then "inverted" to retrieve from the data the values of the other, unobservable variables. This inversion is possible if a one-to-one relationship exists between the model's parameters and the data. When this is so, the model is exactly identified and the equilibrium spatial distribution of activities is unambiguously determined. One feature of quantitative models is worth stressing here: the equilibrium must be unique to ensure that the model-based counterfactuals have unambiguous implications. Yet, settings with increasing returns or externalities typically display several equilibria. In this case, the inverse mapping need not be unique. Uniqueness may be obtained by introducing restrictions on the parameters for the dispersion forces to dominate the agglomeration forces. Alternatively, adding a sufficient amount of heterogeneity to the model often allows uniqueness to be proven (Herrendorf, Valentinyi, and Waldmann 2000). As observed by Redding and Rossi-Hansberg (2017), having a large amount of data may be sufficient to show that the mapping is unique (see, e.g., Ahlfeldt et al. 2015).

At this stage, QSMs are used to make predictions about the impact of policies on the equilibrium spatial distribution of activities, which amounts to undertaking numerically a wide array of large comparative statics exercises that cannot be done analytically. Here, QSMs have several merits that are worth stressing. First, the predictions take into account *all* the general equilibrium effects captured by the model, something difficult to achieve with the reduced-form approach. Second, the model yields *quantitative* predictions. More specifically, a theory model can tell us that a specific shock has a positive impact on some key variables, but the model is often unable to say how strong or how weak the effect is. By contrast, the quantitative model gives us information about the magnitude of the effect, while the reduced-form approach focuses on qualitative relationships between parameters and variables.

Third, undertaking large comparative statics with such models allows one to check the theoretical robustness of the results obtained with the "toy models" considered in sections 3 and 4. Fourth, quantitative spatial models permit new questions to be addressed. For example, we know little about how commuting and trade frictions interact to shape the urban system, an issue that Behrens et al. (2017) explore. As expected, commuting and trade costs matter for the size of cities. What is less expected is that Behrens et al. also find that neither type of friction significantly affects the US city-size distribution. This suggests that changing spatial frictions affect the relative size of cities, but not much their location and rank in the urban hierarchy, as in Desmet and Rossi-Hansberg (2013). In other words, the absolute level of spatial frictions would be less important than their relative levels.

Last, different models are isomorphic to each other (e.g., internal economies of scale and monopolistic competition versus external economies of scale and perfect competition). In this case, the predictions are valid for a family of models, not just for the model used. For example, a setting under perfect competition and the Armington assumption is isomorphic to a Helpman-like model with multiple regions, internal economies of scale, and monopolistic competition (Allen and Arkolakis 2014). However, the latter and an Eaton and Kortum (2002) model, where goods are produced under constant returns and perfect competition, do not generate the same predictions (Redding 2016).

Researchers working with QSMs seem to have a taste for Helpman-like models (1998) where the dispersion force lies in a fixed supply of land at each location. This is probably because this force captures the basic idea that a growing population generates higher urban costs (see section 4). Other researchers stress more differences in amenities and productivity. To illustrate, consider the following simple example. Let $A_r(s) > 0$ be the amenity stock available in region r, which need not be evaluated in the same way across workers who have different skill levels, s. For example, individuals may be heterogeneous in their attitudes toward amenities (see Diamond 2016 and Redding 2016 who describe migration by discrete choice models). Under CES preferences, the indirect utility of an s-worker residing in region r is then as follows:

$$V_r(s) = A_r(s)\frac{w_r}{P_r},$$

where w_r (P_r) is the nominal wage (the CES price index) in region r. In this case, workers are sorted according to their productivity, whereas others will be gathered along their preferences for specific amenities.

5.2.2 *What Do Quantitative Spatial Models Deliver?*

When studying the impact of transport projects, the big advantage of QSMs is that they can take into account all the effects associated with the new infrastructure. Obviously, the connected locations are directly affected, but some of the unconnected locations are affected indirectly because some least-cost routes may go through the new links. Consequently, the new infrastructure leads to a redistribution of labor across regions. A prominent example of the general equilibrium approach can be found in Allen and Arkolakis (2014), who develop a continuous location model for the United States and use a rich, in-depth treatment of transport costs, which are modeled by taking into account all geographic details, including the modal choice. Production is assumed to be perfectly competitive; preferences are CES and the Armington assumption explains why trade occurs. Allen and Arkolakis use their setting to assess the effects of the interstate highway system by recomputing all bilateral transport costs without the interstate highway option. This would reduce total GDP by 1.1 percent to 1.4 percent, which means that the interstate highway system is a productive investment as a whole, even though the network is likely to be far from optimally designed.

Redding (2016) undertakes various large comparative statics experiments by changing the value of transport costs between some location pairs in a grid network. According to Helpman (1998), falling transport costs foster the dispersion of activities. Using a Helpman-like approach, Redding finds that the populations of the largest centers tend to decrease after a transport improvement affecting vertical and horizontal routes of the network, which is reassuring. More importantly, by using a multi-regional general equilibrium model, his analysis allows one to determine which regions are positively or negatively affected by the transport improvement, and by how much the equilibrium regional populations change.

The same quantitative methods may also be used to study how the internal structure of a city, which consists of a discrete set of blocks, is affected by various shocks (Redding and Rossi-Hansberg 2017). Along these lines, the best illustration of what has been achieved is probably the setting developed by Ahlfeldt et al. (2015), which is both tractable and amenable to empirical analysis. These authors consider a fabulous natural experiment, i.e., the fall of the Berlin Wall. From the end of World War II to 1989, the city of Berlin was divided into East Berlin and West Berlin, which could be considered as two local economies separated by prohibitively high transport and communication costs. The division of Berlin affected West Berlin through various channels, including the loss of spillovers generated by the concentration of jobs in the prewar CBD and the drop in the number of employment opportunities in East Berlin. The fall of the Berlin Wall was an unanticipated and sudden shock. Ahlfeldt et al. (2015) use this event as a natural experiment to see how the resulting changes in distance between location pairs have affected the location of activities within West Berlin. More specifically, their aim is to explain both qualitatively and quantitatively the observed changes in city structure, including a relocation of the CBD.

For our purpose, the following results are worth emphasizing. First, both spatial production and consumption externalities are substantial and highly localized. Second, consumption externalities matter more than production externalities (15 percent versus 7 percent). Last, the elasticity of productivity with respect to density within cities is significantly higher than that reported in across-city estimations (7 percent versus 3 percent). Unfortunately, the various spatial externalities used in the paper remain black boxes. In particular, the specifications used do not allow one to discriminate between the different agglomeration economies discussed in section 4, while it is not clear what amenities there really are.

5.2.3 *How to Combine Geography and Growth?*

Since the forces at work in growth and geography models are similar, a solid foundation for cross-fertilization exists between the two fields. However, requiring agents to care about the overall distribution of activities over space and time vastly increases the complexity of the problem. Furthermore, working with aggregates is ill-suited because different sectors obey very different spatio-temporal patterns. Hence, we have to consider a multi-sector economy, which adds an additional degree of difficulty. So far, this is beyond reach. In a pioneering paper, Desmet and Rossi-Hansberg (2014) propose a solution that consists of adding enough structure to the model for the agents to ignore the future paths without affecting the payoffs of their current decisions. The argument involves two main steps.

Land and innovation.—Desmet and Rossi-Hansberg (2012) make a welcome contribution by showing that land is key in firms' decision to invest in new technologies. Consider a one-sector economy where a large number of firms produce a homogeneous good under constant returns using land and labor. Space is given by $X = [0, 1]$. Each firm uses a unit lot size and there is a single firm in each location. Thus, land at

$x \in X$ has the nature of an essential and non-replicable input. Since the marginal productivity of labor is decreasing, a firm's marginal cost increases. As a result, each firm has a well-defined size. A firm's production function is well-behaved and given by $Q = AF(1, L)$ where L is the number of workers and A is a Hicks-neutral shifter that describes the firm's level of technology. Note that innovation cannot be appropriated via patents and is specific to one location. Finally, the cost of acquiring technology A is given by $C(A)$ units of the output.

Since land is assigned to the highest bidder, the equilibrium land rent at x under free entry, is given by

$$R(x) = \max_{A,L}\{p(x)A(x)F(1,L) - w(x)L(x) - p(x)C(A)\},$$

where $p(x)$ and $w(x)$ are the output price and wage at location x. Hence, in a perfectly competitive environment with entry, firms choose to invest in better technologies because the land rent accounts explicitly for the costs they bear to acquire these technologies. To put it differently, by innovating, firms can enhance their bid for land and ensure the best locations. Thus, firms invest in innovation until profits net of the cost of innovation are equal to zero. If land is *not* a production factor, for any new technology A, a firm chooses its labor input to equalize price and marginal cost and, therefore, incurs a loss equal to $p(x)C(A)$. In this case, firms choose to produce using the current technology.

Spatial development.—Desmet and Rossi-Hansberg (2014) use the above idea to build the dynamics of a spatial economy on the basis that firms accumulate knowledge through two different channels, i.e., the firms' decisions to innovate and the spillovers received from firms located nearby.

In every period, labor is freely reallocated between sectors and across space; workers do not consume land. There are two final sectors, manufacturing and services, and two inputs, land and labor. Firms choose how much to produce and how much to invest in technology at each period. Goods and services are traded by incurring iceberg transport costs. When firms innovate, they can secure the benefits of innovation during the current period. Before the next period, the new knowledge is diffused across locations according to an exponential distance-decay function. This spillover is the agglomeration force of the model. As for the dispersion force, it stems from the non-replicability of land, which leads to decreasing returns at the firm level. Since each firm is negligible while the benefits of a current innovation last for only one period and diffuse across locations in the subsequent period, its current decision to innovate has no impact on the stock of knowledge available in the next periods. In other words, the firm's decision to innovate is a static one. As a manufacturing or service firm's production function increases with its available stock of knowledge, there is more innovation in a more productive location. This is the source of local growth. Geography matters because both transport costs and technology diffusion are affected by distance.

At this stage, Desmet and Rossi-Hansberg (2014) calibrate their model by borrowing parameter values from the literature and solving it numerically under CES preferences. The aim is to replicate the evolution of the manufacturing and service sectors in the United States since 1950. Discussing all their results would take us too far off course, so we will limit ourselves to those that are directly related to our purpose. First, the model is able to show that the productivity of the service industry has been catching up since 1995. This is explained by the spatial concentration of service firms and the resulting innovation interactions.

Second, assume that manufacturing firms are spatially concentrated (e.g., in the Northeast and Midwest of the United States), which makes them more productive because the diffusion of knowledge is localized. By contrast, the service firms are dispersed and, hence, less productive. As productivity grows steadily in manufacturing, more and more workers shift to services. When shipping the manufactured good is costly, the service firms choose to locate closer to the manufacturing cluster so their workers have better access to goods. This increasing spatial concentration triggers the technological takeoff of the service firms. In other words, *high transport costs foster the collocation of service and manufacturing firms in adjacent clusters*. As service firms become increasingly concentrated, their productivity rises even further because they keep innovating. As a result, the relative price of goods declines as the manufacturing firms cluster in one area while the growth of the service cluster enhances competition for land. This makes it more expensive for manufacturing firms to innovate, thus giving these firms an incentive to move toward less dense areas (e.g., the South of the United States). A two-region setting is too confined for these various patterns to unfold and is, therefore, unable to capture such a dynamic evolution in firms' locational choices.

Last, we have seen that high transport costs entail static welfare losses through fewer gains in specialization. The previous argument shows that high transport costs generate welfare gains by fostering more innovation through the progressive clustering of the service sector. The numerical analysis undertaken by Desmet and Rossi-Hansberg suggests that the dynamic gains associated with high transport costs offset the static welfare losses. In short, *a dynamic, multi-regional setting may lead to results that differ significantly from those obtained by using a static two-region model*, such as those discussed in section 3.[8]

Despite its appeal, the setting proposed by Desmet and Rossi-Hansberg (2014) has awkward features. For example, the equilibrium is independent of the relative costs of shipping goods and services. This is a bit odd because shipping goods is much cheaper than shipping services, which are often non-tradable. As in Rossi-Hansberg (2005), individuals do not consume land and thus may move to and work in a cluster without affecting the corresponding land values. Otherwise, the price of land would be much higher in the cluster and in nearby locations, which makes the corresponding areas less attractive. In this case, it is not clear that the collocation of the two sectors may emerge.

5.2.4 *Limits of Quantitative Spatial Models*

There is no question that the payoffs of QSMs are high. However, QSMs come at a cost. First, specific functional forms must be assumed. Consequently, the results are conditional on the corresponding functions. In our opinion, one of the main pitfalls of QSMs is precisely the repeated use of the same functional form for preferences, so that we do not know how robust the predictions found in the literature are. More specifically, the CES is almost ubiquitous. Admittedly, the CES is very convenient and has great merits that are too well known to be discussed here. However, the CES is also very peculiar. Indeed, the CES combined with the iceberg cost leads to a broad range of properties that are, unfortunately, difficult to generalize (Parenti, Ushchev, and Thisse 2017). But should we worry about this? It seems so because lowering transport costs intensifies competition and leads to lower prices and a smaller number of firms. Furthermore, a constant elasticity of substitution is a bold assumption that disregards the fact that the

[8] Desmet and Rossi-Hansberg (2015) use a framework similar to the one above to study the effects of climate change on the use of land for agriculture and manufacturing.

entry of new varieties crowds the space of product characteristics and makes varieties closer substitutes (Salop 1979). In this case, the elasticity of substitution increases with the number of varieties, which generates the pro-competitive effects generally associated with entry. The CES model of monopolistic competition is unable to account for these first-order effects.

Second, it is commonplace to use an upper-tier Cobb–Douglas utility nesting CES lower-tier utilities. Under the combination of Cobb–Douglas–CES utilities, the welfare effect of a sectoral shock in the whole economy is confined to the direct welfare effect in the sector. Yet, one expects a positive shock to one sector to trigger a reallocation of budget and labor over all sectors. This puts some severe limitations on the general equilibrium effects.

Last, CES preferences display a very special property: in a one-sector economy, the monopolistically competitive outcome is socially optimal. However, in a multi-industry economy, the optimality property ceases to hold because the allocation of resources across sectors is not optimal, as relative prices across markets are distorted while relative prices within markets are not. But does it matter? The answer is probably yes. Behrens et al. (2016) is a telling example about the possible welfare biases associated with this combination. They consider a multi-sector model of monopolistic competition with CES and constant absolute risk aversion (CARA) preferences, respectively, and assess the welfare loss within and between sectors generated by positive markups. Quantifying the CARA and CES models on French and British data, Behrens et al. (2016) find that there is a substantial aggregate welfare loss of 6 percent to 8 percent under CARA preferences, while the welfare loss is much lower under CES preferences (less than 1 percent).

In short, SQMs rely on settings that have clear theoretical assumptions, which makes this approach highly desirable. A small number of papers that use quasi-natural experiments as an identification strategy suggest that the data do not reject the CES (Redding and Sturm 2008, Donaldson 2018). However, when models are calibrated exactly, a poor choice of functional specifications easily passes unnoticed. Therefore, it is not unreasonable to question the validity of the structural approach when it systematically uses the same specific model. At a time of rapid advances in computational power and numerical methods, robustness checks about functional forms are called for.

6. *Concluding Remarks*

We have surveyed different strands of literature in an effort to show that a few general ideas can be used to answer the questions raised in the introduction, as well as many others. On the way we have, several times, encountered the trade-off between increasing returns and obstacles to the mobility of goods, people, and information, both of which appear under different guises. Nevertheless, despite this common thread, spatial economics is still searching for a general framework that would encompass regional, urban, and transportation economics. Given the complexity of the issue, the search is likely to continue for a long time. This is why we advocate an agnostic attitude.

Simple and stylized settings that can be fully solved analytically are useful in understanding how various effects interact. The merit of such models is that they bring to the fore new effects that stem from the blending of regional and urban economics. In this respect, the work of Akamatsu et al. (2017) looks promising. These authors link different models within a unifying setup that relies on just two basic classes of dispersion forces. This allows them to make predictions about the impact of transport costs in settings that are significantly more general. Quantitative spatial models are able to tackle the same issues from a broader perspective. Unfortunately, general results are hard to prove while the main results obtained in simple settings may not carry over to heterogeneous geographical settings. Among other things, quantitative models can be used to test the relative robustness of results obtained with simple models.

We concur with Holmes (2010) that the same agnostic attitude should prevail in empirical research. Although a structural approach is preferable because it considers several general equilibrium effects, there is still room for reduced-form approaches where key parameters are estimated. After all, some of the structural parameters used in quantitative models are often borrowed from estimations undertaken in other papers. We also need a more systematic confirmation of results. At some point, it will become necessary to explore more systematically the robustness of the conclusions drawn from a long string of empirical papers devoted to agglomeration (dis)economies and to the impact of transport projects.

The mobility of agents is solved under extreme assumptions. A large number of regional and urban economic models assume identical individuals and costless mobility, while transport economics often assumes that agents are fixed. These are signs of a poor understanding of the issue. A more realistic modeling of residential mobility costs would help one to better understand the evolution of spatial patterns. Most contributions also assume that firms and workers move together. However, it is far from being obvious that the mobility of firms and individuals obey the same rules. Hence, studying the steady states of models in which firms and individuals do not react in unison would be a welcome addition to the literature.

It is well known that results obtained in many economic fields under the assumption of identical agents lack robustness. For example, firms' heterogeneity is critical to understanding firms' behavior in the international marketplace and the implications of different policies. The heterogeneity of individuals is probably as important as that of firms. Discrete choice models, which have started being used in quantitative spatial economics, should be given more attention to capture the heterogeneity of individuals. There is also a need to investigate new analytical tools. For example, potential games have already proven very useful in studying congestion in travel flows (Beckmann, McGuire, and Winsten 1956) and the CP model (Oyama 2009). More work along these lines is called for.[9]

It is our contention that spatial economics has reached a sufficiently mature level to trigger cross-fertilization. The work of Moretti (2011) and Zenou (2009), which lies at the interfaces of labor and spatial economics, provides an illuminating example of what could be accomplished by combining fields. In particular, one may wonder why urban and regional economics remain so far from transportation economics. For example, housing and local labor markets are intimately intertwined with the transport markets. What is more, the issue has a strong policy relevance. How to plant the seeds of urban economics in urban planning, and how urban planning may affect what urban economics teaches us, is also critically in short supply.

Transportation economists should pay more attention to the wider benefits associated with the construction of new transport infrastructure, while regional and urban economists should pay more attention to

[9] Researchers appear to have missed the fact that many spatial models have the particular structure of a *population game*, i.e., games with a large number of small and anonymous players (firms or workers), a finite number of pure strategies (the locations), and continuous payoffs. Such games always have a Nash equilibrium in pure strategies and display convenient properties (Sandholm 2010).

the role of transport networks. Spatial economists should no longer treat the transport sector as a black box, but rather as an industrial sector per se that has its own specificities. In addition, the almost ubiquitous iceberg assumption misses important characteristics of transport costs such as density economies and distance economies. Transportation economists also expect regional and urban economists to help them assess the locational and welfare impacts of real-world projects. Examples are numerous and include the provision of new transport infrastructures in the United States, the construction of a large mass transport infrastructure such as the Grand Paris Express, the construction of highways in Western China, and the wide array of transport projects discussed at the European Commission.

A final comment is in order. Economic development and underdevelopment is one of the facets of the lumpy distribution of activities. The material covered in this survey is relevant for economies where institutions, such as land titles, patents, building regulation authorities, work well and governments supply essential infrastructure. This holds for poor or rich countries, for the nineteenth as well as for the twentieth century (Glaeser and Henderson 2017). However, when the economy is mainly informal and market institutions do not function well, we need new models.

References

Aguirregabiria, Victor, and Junichi Suzuki. 2016. "Empirical Games of Market Entry and Spatial Competition in Retail Industries." In *Handbook on the Economics of Retailing and Distribution*, edited by Emek Basker, 201–32. Cheltenham: Edward Elgar.

Ahlfeldt, Gabriel M., Stephen J. Redding, Daniel M. Sturm, and Nikolaus Wolf. 2015. "The Economics of Density: Evidence from the Berlin Wall." *Econometrica* 83 (6): 2127–89.

Akamatsu, Takashi, Tomoya Mori, Minoru Osawa, and Yuki Takayama. 2017. "Spatial Scale of Agglomeration and Dispersion: Theoretical Foundations and Empirical Implications." Research Institute of Economy, Trade and Industry (RIETI) Discussion Paper 17-E-125.

Akamatsu, Takashi, Yuki Takayama, and Kiyohiro Ikeda. 2012. "Spatial Discounting, Fourier, and Racetrack Economy: A Recipe for the Analysis of Spatial Agglomeration Models." *Journal of Economic Dynamics and Control* 36 (11): 1729–59.

Albouy, David. 2009. "The Unequal Geographic Burden of Federal Taxation." *Journal of Political Economy* 117 (4): 635–67.

Albouy, David, Gabriel Ehrlich, and Minchul Shin. 2018. "Metropolitan Land Values." *Review of Economics and Statistics* 100 (3): 454–66.

Allen, Treb, and Costas Arkolakis. 2014. "Trade and the Topography of the Spatial Economy." *Quarterly Journal of Economics* 129 (3): 1085–140.

Alonso, William. 1964. *Location and Land Use: Toward a General Theory of Land Rent*. Cambridge, MA: Harvard University Press.

Anas, Alex, and Robin Lindsey. 2011. "Reducing Urban Road Transportation Externalities: Road Pricing in Theory and in Practice." *Review of Environmental Economics and Policy* 5 (1): 66–88.

Anas, Alex, and Yu Liu. 2007. "A Regional Economy, Land Use, and Transportation Model (RELU-TRAN): Formulation, Algorithm Design, and Testing." *Journal of Regional Science* 47 (3): 415–55.

Anderson, James E. 2011. "The Gravity Model." *Annual Review of Economics* 3: 133–60.

Anderson, James E., and Eric van Wincoop. 2004. "Trade Costs." *Journal of Economic Literature* 42 (3): 691–751.

Anderson, Simon P., Jacob K. Goeree, and Roald Ramer. 1997. "Location, Location, Location." *Journal of Economic Theory* 77 (1): 102–27.

Anderstig, Christer, Svante Berglund, Jonas Eliasson, and Matts Andersson. 2016. "Congestion Charges and Labour Market Imperfections." *Journal of Transport Economics and Policy* 50 (2): 113–31.

Arnott, Richard. 2007. "Congestion Tolling with Agglomeration Externalities." *Journal of Urban Economics* 62 (2): 187–203.

Arnott, Richard, André de Palma, and Robin Lindsey. 1993. "A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand." *American Economic Review* 83 (1): 161–79.

Arnott, Richard, and Kenneth Small. 1994. "The Economics of Traffic Congestion." *American Scientist* 82 (5): 446–55.

Arzaghi, Mohammad, and J. Vernon Henderson. 2008. "Networking off Madison Avenue." *Review of Economic Studies* 75 (4): 1011–38.

Au, Chun-Chung, and J. Vernon Henderson. 2006. "Are Chinese Cities Too Small?" *Review of Economic Studies* 73 (3): 549–76.

Bacolod, Marigee, Bernardo S. Blum, and William C. Strange. 2009. "Skills in the City." *Journal of Urban Economics* 65 (2): 136–53.

Bairoch, Paul. 1988. *Cities and Economic Development: From the Dawn of History to the Present*. Chicago: University of Chicago Press.

Baldwin, Richard, Rikard Forslid, Philippe Martin, Gianmarco Ottaviano, and Frederic Robert-Nicoud. 2003. *Economic Geography and Public Policy*. Princeton: Princeton University Press.

Baum-Snow, Nathaniel. 2007. “Did Highways Cause Suburbanization?” *Quarterly Journal of Economics* 122 (2): 775–805.

Baum-Snow, Nathaniel, J. Vernon Henderson, Matthew A. Turner, Qinghua Zhang, and Loren Brandt. Forthcoming. “Does Investment in National Highways Help or Hurt Hinterland City Growth?” *Journal of Urban Economics*.

Beckmann, Martin J. 1972. “Spatial Cournot Oligopoly.” *Papers in Regional Science* 28 (1): 37–48.

Beckmann, Martin J. 1976. “Spatial Equilibrium in the Dispersed City.” In *Mathematical Land Use Theory*, edited by G. J. Papageorgiou, 117–25. Lexington: Lexington Books.

Beckmann, Martin, C. B. McGuire, and Christopher B. Winsten. 1956. *Studies in the Economics of Transportation*. New Haven: Yale University Press.

Behrens, Kristian, Gilles Duranton, and Frédéric Robert-Nicoud. 2014. “Productive Cities: Sorting, Selection, and Agglomeration.” *Journal of Political Economy* 122 (3): 507–53.

Behrens, Kristian, Carl Gaigné, Gianmarco I. P. Ottaviano, and Jacques-François Thisse. 2007. “Countries, Regions and Trade: On the Welfare Impacts of Economic Integration.” *European Economic Review* 51 (5): 1277–301.

Behrens, Kristian, Carl Gaigné, and Jacques-François Thisse. 2009. “Industry Location and Welfare When Transport Costs Are Endogenous.” *Journal of Urban Economics* 65 (2): 195–208.

Behrens, Kristian, Andrea R. Lamorgese, Gianmarco I. P. Ottaviano, and Takatoshi Tabuchi. 2009. “Beyond the Home Market Effect: Market Size and Specialization in a Multi-country World.” *Journal of International Economics* 79 (2): 259–65.

Behrens, Kristian, Giordano Mion, Yasusada Murata, and Jens Suedekum. 2016. “Distorted Monopolistic Competition.” Düsseldorf Institute for Competition Economics Discussion Paper 237.

Behrens, Kristian, Giordano Mion, Yasusada Murata, and Jens Suedekum. 2017. “Spatial Frictions.” *Journal of Urban Economics* 97: 40–70.

Belenzon, Sharon, and Mark Schankerman. 2013. “Spreading the Word: Geography, Policy, and Knowledge Spillovers.” *Review of Economics and Statistics* 95 (3): 884–903.

Berger, Thor, and Kerstin Enflo. 2017. “Locomotives of Local Growth: The Short- and Long-Term Impact of Railroads in Sweden.” *Journal of Urban Economics* 98: 124–38.

Berliant, Marcus, Shin-Kun Peng, and Ping Wang. 2002. “Production Externalities and Urban Configuration.” *Journal of Economic Theory* 104 (2): 275–303.

Bernard, Andrew B., J. Bradford Jensen, Stephen J. Redding, and Peter K. Schott. 2007. “Firms in International Trade.” *Journal of Economic Perspectives* 21 (3): 105–30.

Berry, Steven, and Joel Waldfogel. 2010. “Product Quality and Market Size.” *Journal of Industrial Economics* 58 (1): 1–31.

Blanchard, Olivier Jean, and Lawrence F. Katz. 1992. “Regional Evolutions.” *Brookings Papers on Economic Activity* 22 (1): 1–61.

Bleakley, Hoyt, and Jeffrey Lin. 2012a. “Portage and Path Dependence.” *Quarterly Journal of Economics* 127 (2): 587–644.

Bleakley, Hoyt, and Jeffrey Lin. 2012b. “Thick-Market Effects and Churning in the Labor Market: Evidence from U.S. Cities.” *Journal of Urban Economics* 72 (2–3): 87–103.

Blonigen, Bruce A. and Anca D. Cristea. 2015. "Air Service and Urban Growth: Evidence from a Quasi-Natural Policy Experiment.” *Journal of Urban Economics* 86 (3): 128–46.

Bonfatti, Roberto, and Steven Poelhekke. 2017. “From Mine to Coast: Transport Infrastructure and the Direction of Trade in Developing Countries.” *Journal of Development Economics* 127: 91–108.

Bosker, Maarten, Steven Brakman, Harry Garretsen, and Marc Schramm. 2012. “Relaxing Hukou: Increased Labor Mobility and China’s Economic Geography.” *Journal of Urban Economics* 72 (2–3): 252–66.

Bosker, Maarten, and Eltjo Buringh. 2017. “City Seeds: Geography and the Origins of the European City System.” *Journal of Urban Economics* 98: 139–57.

Bosquet, Clément, and Henry G. Overman. 2019. “Why Does Birthplace Matter So Much?” *Journal of Urban Economics* 110: 26–34.

Brinkman, Jeffrey C. 2016. “Congestion, Agglomeration, and the Structure of Cities.” *Journal of Urban Economics* 94: 13–31.

Bröcker, Johannes, Artem Korzhenevych, and Carsten Schürmann. 2010. “Assessing Spatial Equity and Efficiency Impacts of Transport Infrastructure Projects.” *Transportation Research Part B: Methodological* 44 (7): 795–811.

Brueckner, Jan K. 2000. “Urban Sprawl: Diagnosis and Remedies.” *International Regional Science Review* 23 (2): 160–71.

Brueckner, Jan K., and Stuart S. Rosenthal. 2009. “Gentrification and Neighborhood Housing Cycles: Will America’s Future Downtowns Be Rich?” *Review of Economics and Statistics* 91 (4): 725–43.

Brueckner, Jan K. and Harris Selod. 2006. "The Political Economy of Urban Transport-System Choice." *Journal of Public Economics* 90 (6–7): 983–1005.

Buzard, Kristy, Gerald A. Carlino, Robert M. Hunt, Jake K. Carr, and Tony E. Smith. 2017. “The Agglomeration of American R&D Labs.” *Journal of Urban Economics* 101: 14–26.

Cairncross, Frances. 2001. *The Death of Distance: How the Communications Revolution Is Changing Our Lives*. Cambridge: Harvard Business Review Press.

Campante, Filipe R., and David Yanagizawa-Drott. 2018. “Long-Range Growth: Economic Development in the Global Network of Air Links.” *Quarterly Journal of Economics* 133 (3): 1395–458.

Campbell, Jeffrey R., and Hugo A. Hopenhayn. 2005. "Market Size Matters." *Journal of Industrial Economics* 53 (1): 1–25.

Chamberlin, Edward. 1933. *The Theory of Monopolistic Competition*. Cambridge: Harvard University Press.

Chandra, Amitabh, and Eric Thompson. 2000. "Does Public Infrastructure Affect Economic Activity? Evidence from the Rural Interstate Highway System." *Regional Science and Urban Economics* 30 (4): 457–90.

Charlot, Sylvie, Carl Gaigné, Frédéric Robert-Nicoud, and Jacques-François Thisse. 2006. "Agglomeration and Welfare: The Core–Periphery Model in the Light of Bentham, Kaldor, and Rawls." *Journal of Public Economics* 90 (1–2): 325–47.

Charnoz, Pauline, Claire Lelarge, and Corentin Trevien. 2018. "Communication Costs and the Internal Organisation of Multi-plant Businesses: Evidence from the Impact of the French High-speed Rail." *Economic Journal* 128 (610): 949–94.

Cheshire, Paul C., Max Nathan, and Henry G. Overman. 2014. *Urban Economics and Urban Policy: Challenging Conventional Policy Wisdom*. Cheltenham and Northampton: Edward Elgar.

Christaller, Walter. 1933. *Die Zentralen Orte in Süd-deutschland*. Jena: Gustav Fischer Verlag.

Christaller, Walter. 1966. *Central Places of Southern Germany*. Englewood Cliffs: Prentice-Hall.

Ciccone, Antonio, and Robert E. Hall. 1996. "Productivity and the Density of Economic Activity." *American Economic Review* 86 (1): 54–70.

Colantone, Italo, and Piero Stanig. 2018. "Global Competition and Brexit." *American Political Science Review* 112 (2): 201–18.

Combes, Pierre-Philippe, Sylvie Démurger, and Shi Li. 2015. "Migration Externalities in Chinese Cities." *European Economic Review* 76: 152–67.

Combes, Pierre-Philippe, Gilles Duranton, and Laurent Gobillon. 2008. "Spatial Wage Disparities: Sorting Matters!" *Journal of Urban Economics* 63 (2): 723–42.

Combes, Pierre-Philippe, Gilles Duranton, and Laurent Gobillon. 2011. "The Identification of Agglomeration Economies." *Journal of Economic Geography* 11 (2): 253–66.

Combes, Pierre-Philippe, Gilles Duranton, Laurent Gobillon, Diego Puga, and Sébastien Roux. 2012. "The Productivity Advantages of Large Cities: Distinguishing Agglomeration from Firm Selection." *Econometrica* 80 (6): 2543–94.

Combes, Pierre-Philippe, and Laurent Gobillon. 2015. "The Empirics of Agglomeration Economies." In *Handbook of Regional and Urban Economics*, Vol. 5, edited by Gilles Duranton, J. Vernon Henderson, and William C. Strange, 247–348. Amsterdam: North-Holland.

Combes, Pierre-Philippe, and Miren Lafourcade. 2005. "Transport Costs: Measures, Determinants, and Regional Policy Implications for France." *Journal of Economic Geography* 5 (5): 319–49.

Costinot, Arnaud, Dave Donaldson, Margaret K. Kyle, and Heidi Williams. 2019. "The More We Die, the More We Sell? A Simple Test of the Home-Market Effect." *Quarterly Journal of Economics* 134 (2): 843–94.

Couture, Victor. 2016. "Valuing the Consumption Benefits of Urban Density." University of California, Berkeley, Memo.

d'Aspremont, Claude, Jean Jaskold Gabszewicz, and Jacques-François Thisse. 1979. "On Hotelling's 'Stability in Competition.'" *Econometrica* 47 (5): 1045–50.

Davis, Donald R., and Jonathan I. Dingel. 2019. "A Spatial Knowledge Economy." *American Economic Review* 109 (1): 153–70.

Davis, Donald R., and David E. Weinstein. 1999. "Economic Geography and Regional Production Structure: An Empirical Investigation." *European Economic Review* 43 (2): 379–407.

Davis, Donald R., and David E. Weinstein. 2002. "Bones, Bombs, and Break Points: The Geography of Economic Activity." *American Economic Review* 92 (5): 1269–89.

Davis, Donald R., and David E. Weinstein. 2003. "Market Access, Economic Geography and Comparative Advantage: An Empirical Assessment." *Journal of International Economics* 59: 1–23.

De Borger, Bruno, and Stef Proost. 2012. "A Political Economy Model of Road Pricing." *Journal of Urban Economics* 71 (1): 79–92.

De Borger, Bruno, and Stef Proost. 2016. "Can We Leave Road Pricing to the Regions? The Role of Institutional Constraints." *Regional Science and Urban Economics* 60: 208–22.

De la Roca, Jorge. 2017. "Selection in Initial and Return Migration: Evidence from Moves across Spanish Cities." *Journal of Urban Economics* 100: 33–53.

De la Roca, Jorge, and Diego Puga. 2017. "Learning by Working in Big Cities." *Review of Economics Studies* 84 (1): 106–42.

De Loecker, Jan, Pinelopi K. Goldberg, Amit K. Khandelwal, and Nina Pavcnik. 2016. "Prices, Markups, and Trade Reform." *Econometrica* 84 (2): 445–510.

de Palma, A., V. Ginsburgh, Y. Y. Papageorgiou, and J.-F. Thisse. 1985. "The Principle of Minimum Differentiation Holds under Sufficient Heterogeneity." *Econometrica* 53 (4): 767–81.

de Rus, Ginés, and Gustavo Nombela. 2007. "Is Investment in High Speed Rail Socially Profitable?" *Journal of Transport Economics and Policy* 41 (1): 3–23.

Desmet, Klaus, and Esteban Rossi-Hansberg. 2012. "Innovation in Space." *American Economic Review: Papers and Proceedings* 102 (3): 447–52.

Desmet, Klaus, and Esteban Rossi-Hansberg. 2013. "Urban Accounting and Welfare." *American Economic Review* 103 (6): 2296–327.

Desmet, Klaus, and Esteban Rossi-Hansberg. 2014. "Spatial Development." *American Economic Review* 104 (4): 1211–43.

Desmet, Klaus, and Esteban Rossi-Hansberg. 2015. "On the Spatial Economic Impact of Global

Warming." *Journal of Urban Economics* 88: 16–37.

Di Addario, Sabrina. 2011. "Job Search in Thick Markets." *Journal of Urban Economics* 69 (3): 303–18.

Diamond, Jared. 1997. *Guns, Germs, and Steel: The Fate of Human Societies*. New York and London: W. W. Norton and Company.

Diamond, Rebecca. 2016. "The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980–2000." *American Economic Review* 106 (3): 479–524.

Dixit, Avinash K., and Joseph E. Stiglitz. 1977. "Monopolistic Competition and Optimum Product Diversity." *American Economic Review* 67 (3): 297–308.

Donaldson, Dave. 2018. "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure." *American Economic Review* 108 (4–5): 899–934.

Donaldson, Dave, and Richard Hornbeck. 2016. "Railroads and American Economic Growth: A 'Market Access' Approach." *Quarterly Journal of Economics* 131 (2): 799–858.

Downs, Anthony. 1957. *An Economic Theory of Democracy*. New York: Harper and Row.

Duranton, Gilles, J. Vernon Henderson, and William C. Strange, eds. 2015. *Handbook of Regional and Urban Economics*. Vol. 5. Amsterdam: North-Holland.

Duranton, Gilles, and Diego Puga. 2004. "Micro-foundations of Urban Agglomeration Economies." In *Handbook of Regional and Urban Economics*, Vol. 4, edited by J. Vernon Henderson and Jacques-François Thisse, 2063–117. Amsterdam: North-Holland.

Duranton, Gilles, and Diego Puga. 2005. "From Sectoral to Functional Urban Specialisation." *Journal of Urban Economics* 57 (2): 343–70.

Duranton, Gilles, and Diego Puga. 2015. "Urban Land Use." In *Handbook of Regional and Urban Economics*, Vol. 5, edited by Gilles Duranton, J. Vernon Henderson, and William C. Strange, 467–560. Amsterdam: North-Holland.

Duranton, Gilles, and Matthew A. Turner. 2011. "The Fundamental Law of Road Congestion: Evidence from US Cities." *American Economic Review* 101 (6): 2616–52.

Duranton, Gilles, and Matthew A. Turner. 2012. "Urban Growth and Transportation." *Review of Economic Studies* 79 (4): 1407–40.

Duranton, Gilles, Peter M. Morrow, and Matthew A. Turner. 2014. "Roads and Trade: Evidence from the US." *Review of Economic Studies* 81 (2): 681–724.

Eaton, Jonathan, and Samuel Kortum. 2002. "Technology, Geography, and Trade." *Econometrica* 70 (5): 1741–79.

Eeckhout, Jan. 2004. "Gibrat's Law for (All) Cities." *American Economic Review* 94 (5): 1429–51.

Eeckhout, Jan, Roberto Pinheiro, and Kurt Schmidheiny. 2014. "Spatial Sorting." *Journal of Political Economy* 122 (3): 554–620.

Eliasson, Jonas, Lars Hultzkranz, Lena Nerhagen, and Lena Smidfelt Rosqvist. 2009. "The Stockholm Congestion-Charging Trial 2006: Overview of Effects." *Transportation Research A: Policy and Practice* 43 (3): 240–50.

Faber, Benjamin. 2014. "Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System." *Review of Economic Studies* 81 (3): 1046–70.

Faggio, Giulia, Olmo Silva, and William C. Strange. 2017. "Heterogeneous Agglomeration." *Review of Economics and Statistics* 99 (1): 80–94.

Feldman, Maryann P., and Dieter F. Kogler. 2010. "Stylized Facts in the Geography of Innovation." In *Handbook of the Economics of Innovation*, Vol. 1, edited by Bronwyn H. Hall and Nathan Rosenberg, 381–410. Amsterdam: North-Holland.

Florida, Richard. 2017. *The New Urban Crisis: How Our Cities Are Increasing Inequality, Deepening Segregation, and Failing the Middle Class—And What We Can Do about It*. New York: Basic Books.

Fujita, Masahisa. 1989. *Urban Economic Theory: Land Use and City Size*. Cambridge: Cambridge University Press.

Fujita, Masahisa, Paul Krugman, and Tomoya Mori. 1999. "On the Evolution of Hierarchical Urban Systems." *European Economic Review* 43 (2): 209–51.

Fujita, Masahisa, Paul Krugman, and Anthony J. Venables. 1999. *The Spatial Economy: Cities, Regions, and International Trade*. Cambridge: MIT Press.

Fujita, Masahisa, and Hideaki Ogawa. 1982. "Multiple Equilibria and Structural Transition of Non-monocentric Urban Configurations." *Regional Science and Urban Economics* 12 (2): 161–96.

Fujita, Masahisa, and Jacques-François Thisse. 2013. *Economics of Agglomeration: Cities, Industrial Location, and Globalization*. 2nd ed. Cambridge: Cambridge University Press.

Gabaix, Xavier. 1999. "Zipf's Law for Cities: An Explanation." *Quarterly Journal of Economics* 114 (3): 739–67.

Gallup, John Luke, Jeffrey D. Sachs, and Andrew D. Mellinger. 1999. "Geography and Economic Development." *International Regional Science Review* 22 (2): 179–232.

Ganong, Peter, and Daniel Shoag. 2017. "Why Has Regional Income Convergence in the U.S. Declined?" *Journal of Urban Economics* 102: 76–90.

Garcia-López, Miquel-Ángel, Adelheid Holl, and Elisabet Viladecans-Marsal. 2015. "Suburbanization and Highways in Spain When the Romans and the Bourbons Still Shape Its Cities." *Journal of Urban Economics* 85: 52–67.

Gibbons, Stephen, Henry G. Overman, and Panu Pelkonen. 2014. "Area Disparities in Britain: Understanding the Contribution of People vs. Place through Variance Decompositions." *Oxford Bulletin of Economics and Statistics* 76 (5): 745–63.

Giroud, Xavier. 2013. "Proximity and Investment: Evidence from Plant-Level Data." *Quarterly Journal of Economics* 128 (2): 861–915.

Glaeser, Edward L. 1999. "Learning in Cities." *Journal of Urban Economics* 46 (2): 254–77.

Glaeser, Edward L. 2011. *Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier*. New York: Penguin

Books.

Glaeser, Edward L., Joseph Gyourko, and Raven E. Saks. 2006. "Urban Growth and Housing Supply." *Journal of Economic Geography* 6 (1): 71–89.

Glaeser, Edward, and J. Vernon Henderson. 2017. "Urban Economics for the Developing World: An Introduction." *Journal of Urban Economics* 98: 1–5.

Glaeser, Edward L., Hedi D. Kallal, José A. Scheinkman, and Andrei Shleifer. 1992. "Growth in Cities." *Journal of Political Economy* 100 (6): 1126–52.

Glaeser, Edward L., Jed Kolko, and Albert Saiz. 2001. "Consumer City." *Journal of Economic Geography* 1 (1): 27–50.

Glaeser, Edward L., and David C. Maré. 2001. "Cities and Skills." *Journal of Labor Economics* 19 (2): 316–42.

Glaeser, Edward L., and Giacomo A. M. Ponzetto. 2018. "The Political Economy of Transportation Investment." *Economics of Transportation* 13: 4–26.

Glaeser, Edward L., and Matthew G. Resseger. 2010. "The Complementarity between Cities and Skills." *Journal of Regional Science* 50 (1): 221–44.

Gobillon, Laurent, and Carine Milcent. 2013. "Spatial Disparities in Hospital Performance." *Journal of Economic Geography* 13 (6): 1013–40.

Gollin, Douglas, Remi Jedwab, and Dietrich Vollrath. 2016. "Urbanization with and without Industrialization." *Journal of Economic Growth* 21 (1): 35–70.

Greenstone, Michael, Richard Hornbeck, and Enrico Moretti. 2010. "Identifying Agglomeration Spillovers: Evidence from Winners and Losers of Large Plant Openings." *Journal of Political Economy* 118 (3): 536–98.

Hall, Jonathan D. 2018. "Pareto Improvements from Lexus Lanes: The Effects of Pricing a Portion of the Lanes on Congested Highways." *Journal of Public Economics* 158: 113–25.

Handbury, Jessie, and David E. Weinstein. 2015. "Goods Prices and Availability in Cities." *Review of Economic Studies* 82 (1): 258–96.

Hanson, Gordon H. 2005. "Market Potential, Increasing Returns, and Geographic Concentration." *Journal of International Economics* 67 (1): 1–24.

Hartwick, John, Urs Schweizer, and Pravin Varaiya. 1976. "Comparative Statics of a Residential Economy with Several Classes." *Journal of Economic Theory* 13 (3): 396–413.

Head, Keith, and Thierry Mayer. 2004. "The Empirics of Agglomeration and Trade." In *Handbook of Regional and Urban Economics*, Vol. 4, edited by J. Vernon Henderson and Jacques-François Thisse, 2609–69. Amsterdam: North-Holland.

Head, Keith, and Thierry Mayer. 2014. "Gravity Equations: Workhorse, Toolkit, and Cookbook." In *Handbook of International Economics,* Vol. 4, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, 131–95. Amsterdam: North-Holland.

Helpman, Elhanan. 1998. "The Size of Regions." In *Topics in Public Economics: Theoretical and Applied Analysis*, edited by David Pines, Efrail Sadka, and Itzhak Zilcha, 33–54. Cambridge: Cambridge University Press.

Helsley, Robert W., and William C. Strange. 1990. "Matching and Agglomeration Economies in a System of Cities." *Regional Science and Urban Economics* 20 (2): 189–212.

Henderson, J. Vernon. 1974. "The Sizes and Types of Cities." *American Economic Review* 64 (4): 640–56.

Henderson, J. Vernon. 1988. *Urban Development: Theory, Fact and Illusion*. Oxford and New York: Oxford University Press.

Henderson, J. Vernon. 2003. "Marshall's Scale Economies." *Journal of Urban Economics* 53 (1): 1–28.

Henderson, J. Vernon, Ari Kuncoro, and Matt Turner. 1995. "Industrial Development in Cities." *Journal of Political Economy* 103 (5): 1067–90.

Henderson, J. Vernon, and Arindam Mitra. 1996. "The New Urban Landscape: Developers and Edge Cities." *Regional Science and Urban Economics* 26 (6): 613–43.

Henderson, J. Vernon, Tim Squires, Adam Storeygard, and David Weil. 2018. "The Global Distribution of Economic Activity: Nature, History, and the Role of Trade." *Quarterly Journal of Economics* 133 (1): 357–406.

Henderson, J. Vernon, and Anthony J. Venables. 2009. "The Dynamics of City Formation." *Review of Economic Dynamics* 12 (2): 233–54.

Herrendorf, Berthold, Ákos Valentinyi, and Robert Waldmann. 2000. "Ruling out Multiplicity and Indeterminacy: The Role of Heterogeneity." *Review of Economic Studies* 67 (2): 295–307.

Hillberry, Russell, and David Hummels. 2008. "Trade Responses to Geographic Frictions: A Decomposition Using Micro-data." *European Economic Review* 52 (3): 527–50.

Hohenberg, Paul M., and Lynn Hollen Lee. 1985. *The Making of Urban Europe, 1000–1950*. Cambridge: Harvard University Press.

Hollander, Stephan, and Arnt Verriest. 2016. "Bridging the Gap: The Design of Bank Loan Contracts and Distance." *Journal of Financial Economics* 119 (2): 399–419.

Holmes, Thomas, J. 2010. "Structural, Experimentalist, and Descriptive Approaches to Empirical Work in Regional Economics." *Journal of Regional Science* 50 (1): 6–22.

Hoover, Edgar Malone. 1948. *The Location of Economic Activity*. New York: McGraw-Hill.

Hornbeck, Richard, and Daniel Keniston. 2017. "Creative Destruction: Barriers to Urban Growth and the Great Boston Fire of 1872." *American Economic Review* 107 (6): 1365–98.

Hotelling, Harold. 1929. "Stability in Competition." *Economic Journal* 39 (153): 41–57.

Hsieh, Chang-Tai, and Enrico Moretti. 2019. "Housing Constraints and Spatial Misallocation." *American Economic Journal: Macroeconomics* 11 (2): 1–39.

Hsu, Wen-Tai. 2012. "Central Place Theory and City Size Distribution." *Economic Journal* 122 (563): 903–32.

Hsu, Wen-Tai, and Hongliang Zhang. 2014. "The

Fundamental Law of Highway Congestion Revisited: Evidence from National Expressways in Japan." *Journal of Urban Economics* 81: 65–76.

Hurter, Arthur P., and Joseph S. Martinich. 1989. *Facility Location and the Theory of Production*. Berlin: Springer.

Ikeda, Kiyohiro, Takashi Akamatsu, and Tatsuhito Kono. 2012. "Spatial Period-Doubling Agglomeration of a Core–Periphery Model with a System of Cities." *Journal of Economic Dynamics and Control* 36 (5): 754–78.

Jedwab, Remi, and Alexander Moradi. 2016. "The Permanent Effects of Transportation Revolutions in Poor Countries: Evidence from Africa." *Review of Economics and Statistics* 98 (2): 268–84.

Kahneman, Daniel, Alan B. Krueger, David A. Schkade, Norbert Schwarz, and Arthur A. Stone. 2004. "A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method." *Science* 306 (5702): 1776–80.

Kaldor, Nicholas. 1935. "Market Imperfection and Excess Capacity." *Economica* 2 (5): 35–50.

Kanemoto, Yoshitsugu. 2013. "Second-best Cost–Benefit Analysis in Monopolistic Competition Models of Urban Agglomeration." *Journal of Urban Economics* 76: 83–92.

Keller, Wolfgang. 2002. "Geographic Localization of International Technology Diffusion." *American Economic Review* 92 (1): 120–42.

Kennan, John, and James R. Walker. 2011. "The Effect of Expected Income on Individual Migration Decisions." *Econometrica* 79 (1): 211–51.

Knight, Brian. 2004. "Parochial Interests and the Centralized Provision of Local Public Goods: Evidence from Congressional Voting on Transportation Projects." *Journal of Public Economics* 88 (3–4): 845–66.

Kok, Suzanne. 2014. "Town and City Jobs: How Your Job Is Different in Another Location." *Regional Science and Urban Economics* 49: 58–67.

Koopmans, Tjalling C., and Martin Beckmann. 1957. "Assignment Problems and the Location of Economic Activities." *Econometrica* 25 (1): 53–76.

Koster, Paul R., and Hans R. A. Koster. 2015. "Commuters' Preferences for Fast and Reliable Travel: A Semi-parametric Estimation Approach." *Transportation Research Part B: Methodological* 81 (1): 289–301.

Krugman, Paul. 1980. "Scale Economies, Product Differentiation, and the Pattern of Trade." *American Economic Review* 70 (5): 950–59.

Krugman, Paul. 1991. "Increasing Returns and Economic Geography." *Journal of Political Economy* 99 (3): 483–99.

Krugman, Paul, and Anthony J. Venables. 1995. "Globalization and the Inequality of Nations." *Quarterly Journal of Economics* 110 (4): 857–80.

Leamer, Edward E. 2007. "A Flat World, a Level Playing Field, a Small World After All, or None of the Above? A Review of Thomas L. Friedman's *The World Is Flat*." *Journal of Economic Literature* 45 (1): 83–126.

Leamer, Edward E., and Michael Storper. 2001. "The Economic Geography of the Internet Age." *Journal of International Business Studies* 32 (4): 641–65.

Lee, Sanghoon, and Jeffrey Lin. 2018. "Natural Amenities, Neighborhood Dynamics, and Persistence in the Spatial Distribution of Income." *Review of Economic Studies* 85 (1): 663–94.

Lin, Yatang. 2017. "Travel Costs and Urban Specialization Patterns: Evidence from China's High Speed Railway System." *Journal of Urban Economics* 98: 98–123.

Lösch, August. 1940. *Die Räumliche Ordnung der Wirtschaft*. Jena: Gustav Fischer.

Lösch, August. 1954. *The Economics of Location*. New Haven: Yale University Press.

Lucas, Robert E., Jr., and Esteban Rossi-Hansberg. 2002. "On the Internal Structure of Cities." *Econometrica* 70 (4): 1445–76.

Lychagin, Sergey, Joris Pinkse, Margaret E. Slade, and John Van Reenen. 2016. "Spillovers in Space: Does Geography Matter?" *Journal of Industrial Economics* 64 (2): 295–335.

Manning, Alan. 2010. "The Plant Size-Place Effect: Agglomeration and Monoposony in Labour Markets." *Journal of Economic Geography* 10 (5): 717–44.

Marshall, Alfred. 1890. *Principles of Economics*. London: Macmillan.

Matsuyama, Kiminori. 2017. "Geographical Advantage: Home Market Effect in a Multi-region World." *Research in Economics* 71 (4): 740–58.

Mills, Edwin S. 1967. "An Aggregative Model of Resource Allocation in a Metropolitan Area." *American Economic Review* 57 (2): 197–210.

Mion, Giordano, and Paolo Naticchioni. 2009. "The Spatial Sorting and Matching of Skills and Firms." *Canadian Journal of Economics* 42 (1): 28–55.

Moretti, Enrico. 2004. "Human Capital Externalities in Cities." In *Handbook of Regional and Urban Economics*, Vol. 4, edited by J. Vernon Henderson and Jacques-François Thisse, 2243–91. Amsterdam: North-Holland.

Moretti, Enrico. 2011. "Local Labor Markets." In *Handbook of Labor Economics*, Vol. 4B, edited by David Card and Orley Ashenfelter, 1237–313. Amsterdam: North-Holland.

Moretti, Enrico. 2012. *The New Geography of Jobs*. New York: Houghton Mifflin Harcourt Publishing.

Mori, Tomoya. 2012. "Increasing Returns in Transportation and the Formation of Hubs." *Journal of Economic Geography* 12 (4): 877–97.

Muth, Richard F. 1969. *Cities and Housing: The Spatial Pattern of Urban Residential Land Use.* Chicago: University of Chicago Press.

Nitsch, Volker. 2005. "Zipf Zipped." *Journal of Urban Economics* 57 (1): 86–100.

Nocke, Volker. 2006. "A Gap for Me: Entrepreneurs and Entry." *Journal of the European Economic Association* 4 (5): 929–55.

Ogawa, Hideaki, and Masahisa Fujita. 1980. "Equilibrium Land Use Patterns in a Nonmonocentric City." *Journal of Regional Science* 20 (4): 455–75.

O'Hara, Donald J. 1977. "Location of Firms within a Square Central Business District." *Journal of Political Economy* 85 (6): 1189–207.

Ohlin, Bertil. 1968. *Interregional and International Trade*. Cambridge: Harvard University Press [1933].

Okubo, Toshihiro, Pierre M. Picard, and Jacques-François Thisse. 2010. "The Spatial Selection of Heterogeneous Firms." *Journal of International Economics* 82 (2): 230–37.

Osawa, Minoru, Takashi Akamatsu, and Yuki Takayama. 2017. "Harris and Wilson (1978) Model Revisited: The Spatial Period-doubling Cascade in an Urban Retail Model." *Journal of Regional Science* 57 (3): 442–66.

Ottaviano, Gianmarco, and Jacques-François Thisse. 2004. "Agglomeration and Economic Geography." In *Handbook of Regional and Urban Economics*, Vol. 4, edited by J. Vernon Henderson and Jacques-François Thisse, 2563–608. Amsterdam: North-Holland.

Oyama, Daisuke. 2009. "History versus Expectations in Economic Geography Reconsidered." *Journal of Economic Dynamics and Control* 33 (2): 394–408.

Parenti, Mathieu, Philip Ushchev, and Jacques-François Thisse. 2017. "Toward a Theory of Monopolistic Competition." *Journal of Economic Theory* 167: 86–115.

Parry, Ian W. H., and Kenneth A. Small. 2009. "Should Urban Transit Subsidies Be Reduced?" *American Economic Review* 99 (3): 700–724.

Parry, Ian W. H., Margaret Walls, and Winston Harrington. 2007. "Automobile Externalities and Policies." *Journal of Economic Literature* 45 (2): 373–99.

Pigou, Arthur C. 1932. *The Economics of Welfare*. 4th ed. London: Macmillan.

Pomeranz, Kenneth. 2000. *The Great Divergence. China, Europe, and the Making of the Modern World Economy*. Princeton and Oxford: Princeton University Press.

Proost, Stef, Fay Dunkerley, Saskia van der Loo, Nicole Adler, Johannes Bröcker, and Artem Korzhenevych. 2014. "Do the Selected Trans European Transport Investments Pass the Cost Benefit Test?" *Transportation* 41 (1): 107–32.

Puga, Diego. 1999. "The Rise and Fall of Regional Inequalities." *European Economic Review* 43 (2): 303–34.

Puga, Diego. 2010. "The Magnitude and Causes of Agglomeration Economies." *Journal of Regional Science* 50 (1): 203–19.

Redding, Stephen J. 2013. "Economic Geography: A Review of the Theoretical and Empirical Literature." In *Palgrave Handbook of International Trade*, edited by Daniel Bernhofen, Rod Falvey, David Greenaway, and Udo Kreickemeier, 497–531. Berlin: Springer.

Redding, Stephen J. 2016. "Goods Trade, Factor Mobility and Welfare." *Journal of International Economics* 101: 148–67.

Redding, Stephen J., and Esteban Rossi-Hansberg. 2017. "Quantitative Spatial Economics." *Annual Review of Economics* 9: 21–58.

Redding, Stephen J., and Daniel M. Sturm. 2008. "The Costs of Remoteness: Evidence from German Division and Reunification." *American Economic Review* 98 (5): 1766–97.

Redding, Stephen J., and Matthew A. Turner. 2015. "Transportation Costs and the Spatial Organization of Economic Activity." In *Handbook of Regional and Urban Economics*, Vol. 5, edited by Gilles Duranton, J. Vernon Henderson, and William C. Strange, 1339–98. Amsterdam: North-Holland.

Redding, Stephen, and Anthony J. Venables. 2004. "Economic Geography and International Inequality." *Journal of International Economics* 62 (1): 53–82.

Roback, Jennifer. 1982. "Wages, Rents, and the Quality of Life." *Journal of Political Economy* 90 (6): 1257–78.

Robert-Nicoud, Frédéric. 2005. "The Structure of Simple 'New Economic Geography' Models (or, on Identical Twins)." *Journal of Economic Geography* 5 (2): 201–34.

Rosenthal, Stuart S., and Stephen L. Ross. 2015. "Change and Persistence in the Economic Status of Neighborhoods and Cities." In *Handbook of Regional and Urban Economics*, Vol. 5, edited by Gilles Duranton, J. Vernon Henderson, and William C. Strange, 1047–120. Amsterdam: North-Holland.

Rosenthal, Stuart S., and William C. Strange. 2004. "Evidence on the Nature and Sources of Agglomeration Economies." In *Handbook of Regional and Urban Economics*, Vol. 4, edited by J. Vernon Henderson and Jacques-François Thisse, 2119–71. Amsterdam: North-Holland.

Rosenthal, Stuart S., and William C. Strange. 2008. "The Attenuation of Human Capital Spillovers." *Journal of Urban Economics* 64 (2): 373–89.

Rossi-Hansberg, Esteban. 2005. "A Spatial Theory of Trade." *American Economic Review* 95 (5): 1464–91.

Rozenfeld, Hernán D., Diego Rybski, Xavier Gabaix, and Hernán A. Makse. 2011. "The Area and Population of Cities: New Insights from a Different Perspective on Cities." *American Economic Review* 101 (5): 2205–25.

Salop, Steven C. 1979. "Monopolistic Competition with Outside Goods." *Bell Journal of Economics* 10 (1): 141–56.

Samuelson, Paul A. 1954. "The Transfer Problem and Transport Costs, II: Analysis of Effects of Trade Impediments." *Economic Journal* 64 (254): 264–89.

Sandholm, William H. 2010. *Population Games and Evolutionary Dynamics*. Cambridge: MIT Press.

Schiff, Nathan. 2015. "Cities and Product Variety: Evidence from Restaurants." *Journal of Economic Geography* 15 (6): 1085–123.

Sidorov, Alexander V., and Evgeny Zhelobodko. 2013. "Agglomeration and Spreading in an Asymmetric World." *Review of Development Economics* 17 (2): 201–19.

Small, Kenneth A. 2012. "Valuation of Travel Time." *Economics of Transportation* 1 (1–2): 2–14.

Small, Kenneth A., and Jia Yan. 2001. "The Value of 'Value Pricing' of Roads: Second-Best Pricing and Product Differentiation." *Journal of Urban*

Economics 49 (2): 310–36.

Small, Kenneth A., Clifford Winston, and Jia Yan. 2005. "Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability." *Econometrica* 73 (4): 1367–82.

Starrett, David. 1978. "Market Allocations of Location Choice in a Model with Free Mobility." *Journal of Economic Theory* 17 (1): 21–37.

Storeygard, Adam. 2016. "Farther on down the Road: Transport Costs, Trade and Urban Growth in Sub-Saharan Africa." *Review of Economic Studies* 83 (3): 1263–95.

Sveikauskas, Leo. 1975. "The Productivity of Cities." *Quarterly Journal of Economics* 89 (3): 393–413.

Syverson, Chad. 2004. "Market Structure and Productivity: A Concrete Example." *Journal of Political Economy* 112 (6): 1181–222.

Tabuchi, Takatoshi. 1998. "Urban Agglomeration and Dispersion: A Synthesis of Alonso and Krugman." *Journal of Urban Economics* 44 (3): 333–51.

Tabuchi, Takatoshi, and Jacques-François Thisse. 2002. "Taste Heterogeneity, Labor Mobility and Economic Geography." *Journal of Development Economics* 69 (1): 155–77.

Tabuchi, Takatoshi, Jacques-François Thisse, and Xiwei Zhu. 2018. "Does Technological Progress Magnify Regional Disparities?" *International Economic Review* 59 (2): 647–63.

Takahashi, Toshiaki, Hajime Takatsuka, and Dao-Zhi Zeng. 2013. "Spatial Inequality, Globalization, and Footloose Capital." *Economic Theory* 53 (1): 213–38.

Takayama, Yuki, and Masao Kuwahara. 2017. "Bottleneck Congestion and Residential Location of Heterogeneous Commuters." *Journal of Urban Economics* 100: 65–79.

Teulings, Coen N., Ioulia V. Ossokina, and Henri L. F. de Groot. 2018. "Land Use, Worker Heterogeneity and Welfare Benefits of Public Goods." *Journal of Urban Economics* 103: 67–82.

Thomas, Isabelle. 2002. *Transportation Networks and the Optimal Location of Human Activities: A Numerical Geography Approach*. Cheltenham: Edward Elgar Publishing.

Toulemonde, Eric. 2006. "Acquisition of Skills, Labor Subsidies, and Agglomeration of Firms." *Journal of Urban Economics* 59 (3): 420–39.

Turner, Matthew A., Andrew Haughwout, and Wilbert van der Klaauw. 2014. "Land Use Regulation and Welfare." *Econometrica* 82 (4): 1341–403.

van den Berg, Vincent, and Erik T. Verhoef. 2011. "Winning or Losing from Dynamic Bottleneck Congestion Pricing? The Distributional Effects of Road Pricing with Heterogeneity in Values of Time and Schedule Delay." *Journal of Public Economics* 95 (7–8): 983–92.

van Ommeren, Jos N., and Eva Gutiérrez-i-Puigarnau. 2011. "Are Workers with a Long Commute Less Productive? An Empirical Analysis of Absenteeism." *Regional Science and Urban Economics* 41 (1): 1–8.

Vickrey, William S. 1964. *Microstatics*. San Diego: Harcourt, Brace and World.

Vickrey, William S. 1969. "Congestion Theory and Transport Investment." *American Economic Review: Papers and Proceedings* 59 (2): 251–60.

von Thünen, Johann Heinrich. 1966. *Der Isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. Hamburg: Perthes. [1826].

von Thünen, Johann Heinrich. 1966. *The Isolated State*. Oxford: Pergammon Press.

Winston, Clifford. 2013. "On the Performance of the U.S. Transportation System: Caution Ahead." *Journal of Economic Literature* 51 (3): 773–824.

Zenou, Yves. 2009. *Urban Labor Economics*. Cambridge: Cambridge University Press.