NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Stef Proost,
Jacques-Francois Thisse

WHAT CAN BE LEARNED FROM SPATIAL ECONOMICS?

BASIC RESEARCH PROGRAM
WORKING PAPERS

SERIES: ECONOMICS
WP BRP 167/EC/2017

Stef Proost[1] Jacques-Francois Thisse[2]

What Can Be Learned from Spatial Economics?

Abstract

Spatial economics aims to explain the location of economic activity. While the importance of the proximity to natural resources has declined considerably, distance and location have not disappeared from economic life. Recent work in spatial economics indicates that new forces, hitherto outweighed by natural factors, are shaping an economic landscape that, with its many barriers and large inequalities, is anything but flat. The location of economic activity is the outcome of a trade-off between different types of scale economies and costs generated by the transfer of people, goods, and information. This trade-off is used as a guide in our survey of the main developments in regional and urban economics, which refer to different spatial scales. The role of transport is discussed for each subfield. We briefly survey the ingredients that could be useful for a synthesis of regional and urban economics and conclude with general policy insights.

Keywords: location, region, city, transport, land, agglomeration.

JEL Classification: F12; F20; F61; L13; R12; R14

[1]Department of Economics, KULeuven. stef.proost@kuleuven.be
[2]CORE-UCLouvain, NRU Higher School of Economics and CEPR. jacques.thisse@uclouvain.be

Introduction

Spatial economics deals with bringing *location*, *transport*, and *land* into economics. These three concepts are closely intertwined and gathered in the category R of the JEL Classification System. Spatial economics aims to explain where economic activities are located in very distinct real-world situations, which run from the global to the local through the national and the urban. Its main challenge is to explain why there are peaks and troughs in the spatial distribution of wealth and people. Thus, the main task of spatial economics is to identify the microeconomic underpinnings of two types of forces: *centripetal* forces, which lead to concentration of economic activities, and *centrifugal* forces, which induce economic activities to move away. Other, but related, key issues include the prevalence of spatial inequality inside cities, within and between countries despite numerous government actions. Another issue is the remarkable resilience of some places in their economic development.

The relationship between spatial economics and economic theory is complex because spatial economics is fraught with most of the difficulties encountered in economic theory (externalities, increasing returns and imperfect competition). It is, therefore, hardly a shock that spatial economics is both at the core and the periphery of economics. It is at the core in that economic growth has always been and still is geographically uneven, while economic historians have convincingly argued that cities play a fundamental role in the process of economic and social development (Bairoch, 1988; Hohenberg and Lee, 1993). Therefore, space and its parent concepts should be an important component of the tool box of economists. In spite of this, spatial economics and its constituent subfields, such as regional economics, called nowadays economic geography, urban and transportation economics, remain on the periphery, and are hardly represented in graduate programs. We will highlight some possible reasons for this lack of interest, which has lasted until the 1990s.

Since the emergence of a knowledge-based economy, traditional location factors have been replaced with new drivers of regional and urban growth, which rely on human capital and cognitive skills. Within a growing number of countries, a few large cities produce a growing share of their gross domestic product (henceforth GDP), despite very high housing and commuting costs. For example, twenty U.S. metropolitan areas produce about 50% of the American GDP, but accommodate approximately one third of the population. Greater Paris, which accounts for 2% of continental France and 19% of its population, produces more than 30% of the GDP of France. Equally surprising, one of the most robust empirical facts in economics is the Gravity Law, which Head and Mayer (2014) summarize as follows: "Holding constant the product of two countries' sizes, their bilateral trade will, on average, be inversely proportional to the distance between them." Even though we lack a comparable study of migration flows, the few existing contributions conducted for specific countries and/or professions suggest that people are sticky, while migration is also gravitational in nature.

These facts and many others run against the common belief that we live in a world where the tyranny of distance, which has been such a powerful force in human history, no longer exists. But–and it is a big but–despite spectacular drops in communication and transport costs, distance and location have not disappeared from economic life. Contrary to widespread beliefs, we do not live in a world in which the playground is levelled. Recent work in spatial economics shows that new forces, hitherto outweighed by natural factors, are shaping an economic landscape that, with its many barriers and large inequalities, is anything but flat. Transportation cuts across economic geography and urban economics, but in a different way. In the regional context, transportation consists of interregional and international freight trips of inputs and outputs, as well as passenger trips (business trips and leisure trips). In the urban context, the focus is mainly on commuting by means of different transportation modes. In spite of its importance for the subject matter, the literature in spatial economics has paid too little attention to what has been accomplished in transportation economics (and vice versa).

Trading commodities is spatial by its very nature. Until recently, trade theory has focused only on a fairly strange geography in which countries are close enough for the cost of shipping goods internationally to be zero, but far enough apart that no workers or capital-owners can find their way from one country to another (Leamer, 2007). This research strategy is especially surprising since Ohlin (1933, 1968, 97) has 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." Eventually, the Gravity Law has led trade economists to recognize, although belatedly, that free trade does not mean costless trade.

Housing and transport, which are the quintessentially space-related commodities, rank first and second in household expenditure. In the United States, the average expenditure share on housing is 24%, while it is equal to 27% in France. Spending on transport amounts to 17% in the United States and 13.5% in France. Per year, the opportunity cost of time spent in commuting accounts for three to six weeks of work for a Manhattanite and, on average, four weeks of work for a resident of Greater Paris. Space is the substratum of human affairs, but space is also a (consumption and production) good under the form of *land*. The worldwide supply of land vastly exceeds the demand for land. As a consequence, the price of land should be zero. Yet, housing costs vary enormously with the size and composition of cities for reasons that do not depend on the quality of the housing structure. Therefore, the price of land must reflect the scarcity of "something" that differs from land per se.

Complexity renders futile any attempt to seek *the* model of agglomeration. Therefore, economists and regional scientists have developed different theories to account for the existence of a variety of spatial clusters. Evidently, the forces at work at different geographical scales are not the same. However, differences are less pronounced than what it 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. Section 2 highlights the main economic forces that shape the space-economy at different scales. In particular, we will explore a question that has generated endless discussions, that is, how to describe best the process of competition across space? To answer this question, we will borrow tools from general equilibrium theory and industrial organization. In passing, this will allow us to highlight why spatial economics has been at the periphery of economics for so long.

Too often, economists use interchangeably different, yet equally unclear, words such as locations, regions or places without being aware that they often correspond to different spatial units. In doing so, they run the risk of drawing implications that are valid at a certain level of spatial aggregation but not at another. Regional economics and urban economics consider different geographical scales. Regional economics, because its main focus is on very large areas, has neglected land. It may thus be viewed as spatial economics *without* land. By contrast, urban economics is spatial economics *with* land because the land market is critical in the working of cities. Transport being all over the place, we may view spatial economics as the all-encompassing field.

Section 3 focuses on the so-called *regional question*, that is, the existence of sizeable and lasting disparities in GDP per capita within nations or regional trade blocks such as the European Union. What distinguishes regional economics from trade theory is the mobility of production factors. Even though speculation on regional disparities have never been in short supply, no one before Krugman (1991) had been able to show how they can arise at a stable market equilibrium. We follow in Krugman's footsteps and adopt the "new economic geography" approach to regional economics. Urbanization being the most extreme form of spatial inequality, we move on in Section 4 by studying cities. We deal with the following questions: *why do cities exist*? Why are workers better paid in bigger cities? And why is housing more expensive in large than in small cities? In the wake of its founding fathers (Alonso, 1964; Mills, 1967; Muth, 1969), urban economics has 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 urban economics is to emphasize the importance of non-tradables, such as housing and services produced within the city for the residents, whereas regional economics remains in the tradition of trade theory by focusing on tradables. More recently, the interest has shifted toward the reasons that pertain to the agglomeration of firms and households in a relatively small number of cities. In this approach, cities are not just the containers of economic activities, they are a key player in the social fabric. In Section 5, we discuss various insights for the development of a synthesis of regional and urban economics, while Section 6 concludes with a discussion of some policy issues.

Although many ideas and concepts have been around for a long time, they remain fairly disparate and in need of a synthesis. Nowadays, the state of the art is sufficiently advanced to sketch a unified framework that could be the future backbone of spatial economics. Furthermore, the field has reached a level of maturity that should make it appealing to scholars working in local public finance, labor and environmental economics, to mention a few. Although the on-going research focuses mainly on the empirics of agglomeration, our emphasis will be on theory, while our concluding remarks draw a few policy recommendations. The recent fifth volume of the *Handbook of Regional and Urban Economics* contains several insightful chapters describing the state of the art in the empirical literature. By contrast, many ideas and results obtained in spatial economic theory have been forgotten or keep being rediscovered under different disguises.

What makes spatial economics different?

The fundamental trade-off of spatial economics

As observed by Koopmans (1957) 60 years ago, increasing returns are critical to understand how the space-economy is shaped. To meet this challenge, we appeal to *the fundamental trade-off between increasing returns and transport costs*, which is valid at all spatial scales (Fujita and Thisse, 2013). The intuition is easy to grasp. In the absence of increasing returns, firms would be able to spread their production over an arbitrary 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, firms would concentrate their production within a few giant plants to benefit from the highest possible level of efficiency. Such nonsensical consequences confirm the relevance and importance of the above trade-off. It has been rediscovered several times, and goes back at least to Lösch (1940).

Location matters

The most natural way to think of increasing returns is when a plant with a minimum capacity has to be built before starting production. 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 under the form of a facility designed to supply collective services to consumers. Many public facilities operate under internal increasing returns because the congregation of a large number of people facilitates the provision of collective services that could not be obtained in isolation. Increasing returns may also materialize under a very different form, in which 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:

The first principle of spatial economics: *While many activities can be located almost anywhere, few activities are located everywhere.*

It is in this sense that location matters: though a large number of activities become "footloose", in many countries, many large areas account for no or little economic activity. Indeed, one should not infer from the low value of 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 only seems a paradox: inexpensive shipping of goods makes competition tougher, and thus firms care more about small advantages than they did in a world in which they were protected by the barriers of high transport costs.

Moving goods and people remains costly

The great many estimations of the gravity equation show that distance remains a strong impediment to trade and exchange (Head and Mayer, 2014). Anderson and van Wincoop (2004) estimate that, for developed countries, average trade costs represent 170% of the producer price of manufactured goods. This is quite a high rate for a world in which distance is supposedly disappearing from economic life.

Being governed by a broad range of economic, intangible, time-persistent factors, migration is also sluggish. For example, the estimations undertaken by Tombe and Zhu (2017) suggest that, in China, the average cost of intra-provincial migrations is around 51% of annual income, whereas the average cost of inter-provincial migration ranges from 94 to 98% of annual income in 2000. Further evidence of the low mobility of workers is provided by Bosquet and Overman (2016). Using the British Household Panel Survey from 1991 to 2009, these authors observe that 43.7% of workers worked only in the area where they were born. Such a strong empirical evidence makes it hard to believe that migration costs are a second-order force.

That both the transport of goods and people should remain costly despite the Internet and other new communication devices has a major implication for the organization of the space-economy:

The second principle of spatial economics: *The world is not flat because what happens near us matters more than what happens far from us.*

Combining the first and second principles of spatial economics has led us to formulate what we see as the fundamental trade-off of spatial economics:

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

We may already conclude that high (resp., low) transport costs promote the dispersion (resp., agglomeration) of economic activities, while strong (resp., weak) increasing returns act as a centripetal (resp., centrifugal) force. This trade-off is valid on all spatial scales (cities, regions, countries, and continents), which makes it a valuable analytical tool.[3] For example, drops in commuting costs have allowed cities to grow, while cheaper haulage has permitted the supply of remote markets by large firms. We will return to this in the next two sections.

[3]Observe that the same trade-off is key in the literature on multinational enterprises (Antràs and Yeaple, 2014).

Space and general equilibrium

The history of the relationship between spatial economics and general equilibrium theory is both complex and obscure. It is complex because it is fraught with difficulties that have been put aside for simplicity. It is obscure because the several attempts made over the last 50 years to clarify this relationship have befuddled the debate with confusing answers. Yet, the argument is beautifully simple. Assume a competitive equilibrium in which firms and households have chosen their locations and trade goods whose transport is costly. By the First Theorem of Welfare Economics, if a competitive equilibrium exists, it is Pareto optimal. Therefore, at the prevailing market prices the aggregate production and transport costs $C + T$ within the economy must be minimized. As a consequence, it must be that

$$C_{q_i}(q_1, ..., q_n) + T_{q_i}(q_1, ..., q_n) = 0 \qquad i = 1, ..., n,$$

where n is the number of goods. Since transport costs increase with the quantity shipped of good i ($T_{q_i} > 0$), firms must operate at a scale where their average cost decreases, that is, $C_{q_i} < 0$. Yet, a competitive equilibrium cannot involve profit-maximizing firms operating at a decreasing average cost. For C_{q_i} to be non-negative, $T_{q_i} = 0$ must hold, that is, backyard capitalism must prevail. As a consequence, there exists no competitive equilibrium with trade. This argument has been generalized by Starrett (1978) as follows.

The spatial impossibility theorem. *Consider an economy endowed with a finite number of locations. If a consumer's utility function is independent of where the consumer resides, a firm's production function is independent of the firm's location, and shipping goods between locations is costly, then there exists no competitive equilibrium involving the shipment of goods across locations.*

Thus, if a competitive equilibrium exists, the spatial impossibility theorem implies that each location operates as an autarky. If there is no ex ante reason for economic agents to distinguish between locations and if activities are perfectly divisible, each agent can operate at an arbitrarily small level: firms and consumers succeed in reducing transport costs 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. What the spatial impossibility theorem says is 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.

One solution to obviate the implications of the spatial impossibility theorem consists in assuming that locations are differentiated by Ricardian attributes. Though analytically convenient, this approach does not help us to understand why some places are a priori more productive than others. Even though there might be just a few suitable places to host particular production activities, spatial inhomogeneities may be useful to explain *where* cities arise, but not *why* they exist. Relying only on Ricardian advantages to explain the existence of large urban agglomerations and sizable trade flows amounts to playing Hamlet without the Prince.

Given the role played by the competitive model in economic theory, there can be little doubt that the spatial impossibility theorem, together with the need to account for scale economies, acted as an impediment to the development of spatial economics until the monopolistic competition revolution of the 1980s. Incidentally, the renewed interest in agglomeration economies brought about by the influential work of Glaeser *et al.* (1992) is concomitant to Krugman's (1991) introduction of monopolistic competition in spatial economics.

Notwithstanding the possible advantages of using exogenous locational inhomogeneities, the spatial impossibility theorem has a major implication for spatial economics. To study the existence of a wide range of economic agglomerations, we have to consider *scale economies* (or size effects), *spatial externalities* (also called spillovers) or *imperfect competition* (oligopolistic or monopolistic). The choice of a particular modeling strategy depends on the spatial scale under consideration. Because their extent is often geographically limited, spatial externalities are mainly relevant for studying issues arising on a local scale (see Section 4). The advantage of using spatial externalities is that they are consistent with the competitive setup. Because it focuses on issues arising on a global scale where spatial externalities are likely to be absent, regional economics relies on the combination internal increasing returns and monopolistic competition to capture the *pecuniary externalities* generated by the mobility of production factors, whereas urban economics abides to *technological externalities* (see Section 3). Before proceeding, we must discuss a third approach to competition in space, which is rooted in industrial organization.

Spatial competition

In his review of Chamberlin's (1933) book, Kaldor (1935), but also Lösch (1940), argued that consumers' dispersion molds competition across space in a very specific way: whatever the total number of firms in the industry, each one competes more vigorously with its immediate neighbors than with more distant firms. Such a market description, which became to be known as *spatial competition*, departs radically from the perfectly competitive setting. The argument goes as follows. Because consumers are spatially dispersed, they differ in their access to the same firm. As a consequence, 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 on the consumers situated in its vicinity, which enables firms to choose their own prices in a strategic environment.

The earliest fully analytical analysis of this problem is due to Vickrey (1964) and Beckmann (1972) who studied how homogeneous firms equidistantly distributed over a one-dimensional space, whence firms have an address, compete to attract consumers who are evenly distributed over the same line. Each consumer buys one unit of the good, while travelling costs are proportional in distance. Firm i has only two neighbors located at a 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 the neighboring firms. In other words, their isolation avails them only local monopoly power, for firm i's demand depends upon the prices set by the neighboring firms:

$$x_i(p_{i-1}, p_i, p_{i+1}) = \max\left\{0, \frac{p_{i-1} - 2p_i + p_{i+1} + 2t\Delta}{2t}\right\}.$$

Consequently, only the behavior of the neighboring firms counts for firm i. Since a reduction in firm i's price significantly affects the demands of the two neighboring firms, they will 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 $i+2$ Pursuing this line of reasoning implies that firm i's action has consequences that spread out in two directions. However, the amplitude of the effect decreases with the distance from firm i, similarly to how a wave spreads out on the surface of a pond after a rock has been thrown into it. This market structure is by nature *oligopolistic* in that each firm is only concerned directly with a small number of competitors. 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, as the travel rate t, the inter-firm distance Δ, or both decrease, firms charge a lower price because competition gets 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.

But how do firms choose where to produce under spatial competition? Hotelling (1929) proposed a solution in a path-breaking study in which two sellers choose, first, where to set up their stores along Main Street, which is described by a linear segment, and, then, the price at which they supply their customers. Whereas the individual purchase decision is discontinuous – a consumer buying only from one firm – firms' aggregated demands are continuous with respect to prices. 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 agglomerate at the market center. If true, this provides us with a rationale for the clustering of firms supplying spatially dispersed consumers. Unfortunately, Hotelling's analysis was plagued by a mistake that invalidates his main conclusion: when firms are sufficiently close, the corresponding subgame does not have a Nash equilibrium in pure strategies. This negative conclusion has led d'Aspremont *et al.* (1979) to modify the Hotelling setting by assuming that consumers' travel 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 *et al.* 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, a great many number of contributions in industrial organization have revisited this setup. The main message is simple: the market center is the most attractive place along Main Street if the duopolists sell goods that are sufficiently differentiated in vertically - quality - or horizontally - variety attributes. When travel costs are low, being geographically isolated is no longer a protection against competition. As a consequence, firms choose to relax competition by being differentiated in a Lancasterian space of attributes while they locate back-to-back at the center of the city space. In brief, *product differentiation is substituted for geographical dispersion.*

Despite its appeal, which allows one to discriminate between firms' flexibility in changing location or price, the two-stage game approach has difficulties to handle a setting with several firms. In addition, when consumers like variety, market areas overlap, which implies that firms no longer have a natural hinterland but are able to retain customers from very different market segments. Given that individual demands are perfectly inelastic (up to the reservation price), de Palma *et al.* (1985) propose to model consumers' shopping behavior by means of the *multinomial logit*, which is a close relative of CES preferences that occupy center stage in the trade literature. More specifically, if $n-1$ firms are located together at $y \in [0,l]$ and post the same price p, the share of purchases from firm i located at $y_i \in [0,l]$ and selling at price p_i made by consumers residing at $x \in [0,l]$ is given by the following expression:

$$s_i(x) = \frac{\exp - \left(p_i + t\left|x - y_i\right|\right)/\alpha}{\exp - \left(p_i + t\left|x - y_j\right|\right)/\alpha + (n-1)\exp - \left(p + t\left|x - y\right|\right)/\alpha} > 0, \qquad (1)$$

where the parameter $\alpha > 0$ measures the degree of differentiation among varieties. The cluster becomes increasingly attractive as the number of stores, the distance between the isolated firm and the cluster shrinks, or both, which agrees with the Gravity Law. As firm i's market demand is obtained by summing (1) across locations, firms' profits are easily determined, allowing de Palma *et al.* to show the following result:

The geographical clustering of firms. *If* $\alpha t \leq 2$, *then* $p_i^* = c + n/\alpha(n-1)$ *and* $y_i^* = l/2$, *for* $i = 1, ..., n$, *is a Nash equilibrium of the simultaneous game.*

In other words, when the preference for variety is sufficiently strong (α is high enough), travel costs are sufficiently low (t is small enough), or both, firms choose to locate at the market center and to price above marginal cost. We will see in the next section that many regional economic models lead to a similar prediction.

Spatial competition models are appealing because they are well suited for studying new aspects of the market process in space. Unfortunately, they rely on fairly specific assumptions. When they are generalized or cast in a general equilibrium framework, they become very quickly intractable. In particular, showing the existence of a Nash equilibrium turns out to be very problematic. Consequently, 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 externalities. This is what we will do in Sections 3 and 4.

The regional question: a spatial interaction approach

Broadly defined, spatial interaction refers to a wide array of flows subject to various types of spatial frictions. Countries and regions are affected by the growing mobility of commodities, but also by that of production factors. A shock on transport or trade costs affects firms' and workers' incentives when they choose where to locate. It is crucial to have a good understanding of how firms and workers react in order to assess the full impact of trade and transport policies. In this section, we will first discuss how the interregional distribution of activities varies with transport costs, as well as on a few other forces that have not been much studied. We build on the home market effect and the core–periphery model. Next, we will 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.

What drives regional disparities?

Standard theory predicts a market outcome in which production factors receive the same reward regardless of where they operate. When each region is endowed with the same production function that exhibits constant returns to scale as well as a decreasing marginal productivity, capital, or labor, responds to market disequilibrium by moving from regions where it is abundant relative to labor 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 capital and labor in equilibrium would also be the same everywhere due to the equalization of capital-labor ratios.

However, we are far from seeing such a featureless world. To solve this contradiction, new economic geography take 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) CES model of monopolistic competition and the iceberg transport technology. In the former, a large number (formally a continuum) of firms produce a differentiated good, which is sold to consumers who have a preference for variety. Each variety is produced by a single firm and each firm produces a single variety using a fixed and constant marginal requirement of labor. In the latter, when one unit of a variety is moved between regions A and B, only a fraction $1/\tau < 1$ reaches the destination, the missing share $(\tau - 1)/\tau$ having 'melted' on the way. This ingenious modeling trick allows one to integrate positive shipping costs without having to deal explicitly with a transport sector. Nevertheless, we will see in section 3.3 that using the iceberg cost is not an innocuous assumption.

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; the share of workers located in region A is equal to $\theta > 1/2$. Capital is mobile between regions and capital owners seek the higher rate of return; the share λ of capital located 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 that allows them to better exploit scale economies, while the dispersion force is generated by firms' desire to avoid tough competition of the product and labor markets (Fujita *et al.*, 1999; Baldwin *et al.*, 2003). This adds competition to the fundamental trade-off of spatial economics discussed in Section 2, and which is known as 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 operating profits made in each market. Since operating profits are redistributed to capital-owners, their investment decisions, whence their incomes, also affect the spatial distribution of demand, which influences the location of firms.

This system of push and pull 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 they are able to produce at a lower average cost and to supply a bigger pool of consumers (Krugman, 1980; Martin and Rogers, 1995). However, the intensity of competition in the larger region keeps some firms in the smaller region. This result that has been coined the *home market effect* (henceforth HME). Due to its size advantage, one expects the larger region to attract more firms than the other. 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.

It is empirically well documented that firms vastly differ in productivity. Importantly, the HME still holds when firms are cost-heterogeneous (Nocke, 2006). In this case, the gathering of the more productive firms renders competition tougher in the larger region, which leads the inefficient firms to locate far apart to avoid the devastating effects of competition with efficient firms. In other words, the spatial selection of firms sparks a productivity gap between regions. Using U.S. data on the concrete industry, Syverson (2004) observes that inefficient firms barely survive in large competitive markets and tend to leave them.

How does a lowering of interregional transport costs affect this result? At first glance, one could expect the proximity effect to be weaker when transport costs are lower. In fact, the opposite holds true: more firms choose to set up in the larger region when it gets cheaper to ship goods between the two regions. This somewhat paradoxical result can be understood as follows. On the one hand, lower transport costs make exporting to the smaller market easier; on the other hand, 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 de-industrialized 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 induce firms to pull out of the smaller region. This result may come as a surprise to those who forget that highways run both ways.

Nevertheless, Takahashi *et al.* (2013) show that, as transport costs steadily decrease, both the equilibrium wage and manufacturing share, first, rise because workers enjoy higher incomes and, then, fall because competition in the labor market yields high labor costs. This suggests a *bell-shaped curve of spatial development*, which we will encounter again below. Accordingly, if lowering transport costs initially fosters a more intensive agglomeration of firms, its continuation is liable to generate a redeployment of activities that could lead to a kind of geographical evening-out.

Unfortunately, the HME cannot be readily extended to multi-regional set-ups because there is no obvious benchmark against which to measure the "more than proportionate" share of firms (Behrens *et al.*, 2009). The new fundamental ingredient that a multi-regional 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. On the contrary, when there are more than two regions, any global or local change in the transport network, such as the construction of a major transportation link, is likely to trigger complex effects that vary in non-trivial ways with the properties of the graph representing the transportation network. Yet, a few suggestive results may be obtained. For example, a local reduction in transport costs between two adjacent regions leads to an increase in their joint GDP. However, if one region is sufficiently larger than the other, the former attracts firms at the expense of the latter (Behrens *et al.*, 2007). A possible way out to the dimensionality problem is the use of infinitely many regions (Rossi-Hansberg, 2004). In this case, what happens in a region has no impact on the others.

Given the foregoing, it is no surprise that the empirical evidence regarding the HME is mixed (Davis and Weinstein, 2003; Head and Mayer, 2004). Intuitively, however, it is reasonable to expect the forces highlighted by the HME to be at work in many real-world situations. But how can we check this? There are two possible ways out. First, although it is hard to test the HME, there is a wealth of evidence suggesting 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. Redding and Sturm (2008) exploited the political division of Germany as a natural experiment to show how the loss of market access for cities in West Germany located close to the border made these cities grow much less. After a careful review of the state of the art, Redding (2011) concludes that "there is not only an association but also a causal relationship between market access and the spatial distribution of economic activity."

Second, since there is no hope of deriving general results for multi-regional economies, it is reasonable to try to solve numerically spatial general equilibrium models where transportation networks for actual and randomly selected networks. For this, one needs a mathematical framework that is tractable but yet rich enough to analyze meaningful effects. This is what several authors aim to accomplish in the last wave of research that we will discuss in Section 5.

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.

Can a core-periphery structure be a stable equilibrium outcome?

One of the most natural ways to think of an 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 has 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 remains otherwise symmetric.

The workhorse of the core-periphery (henceforth CP) model is again the Dixit-Stiglitz-iceberg model, which implies that the total number of firms is determined by the number of workers. 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 of Krugman's paper that dwells on pecuniary external effects. 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 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. Such effects are of particular importance in imperfectly competitive markets as prices fail to reflect the true social value of individual decisions. Hence, studying the full impact of migration requires a full-fledged general equilibrium framework, which captures not only the interactions between product and labor markets, but also the double role played by individuals as both workers and consumers.

Two main effects are at work: one involving firms and the other workers. First, if a region happens to have a higher demand for tradable goods, it hosts a more than proportionate share of manufacturing firms, which pushes nominal wages up. Second, the presence of more firms means a wide range of local varieties, whence a lower local price index – a cost-of-living effect. Accordingly, the real wage increases, and this 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 becomes the periphery.

Even though this process seems to generate a "snowball" effect, it is not so clear that it will always develop according to the foregoing prediction. Indeed, the argument above ignores several key impacts of migration on the labor market. On the one hand, the increased supply of labor in the region of destination will tend to push nominal wages down. On the other hand, the increase in local demand for tradable goods leads to a higher demand for labor. The final impact on nominal wages is thus hard to predict. Likewise, there is increased competition in the product market, which everything else equal reduces firms' profits. The combination of all these effects may spark a "snowball meltdown", which may result in the spatial dispersion of firms and workers.

Krugman (1991) adds a genuine dispersion force to his setup by considering a second sector, e.g. agriculture. 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 incites firms to choose different locations because they enjoy a proximity advantage in supplying the local farmers.

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, interregional shipments of goods are 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 will concentrate into the core. In this way, firms are able to exploit increasing returns by selling more goods in the regions benefiting from the market expansion effects sparked by the migration of workers without losing much business in the smaller markets. Thus, the mobility of labor may exacerbate the HME, the reason being that the size of local markets changes with labor migration. The CP model, therefore, allows for the possibility of *convergence* or *divergence* between regions. Krugman's work appealed because regional disparities emerge as a *stable equilibrium* that is the unintentional consequence of decisions made by a large number of economic agents pursuing their own interests.

To sum up, we have:

The core-periphery structure. *Assume that workers are mobile. When transport costs are sufficiently low, the manufacturing 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 can make some places bigger than others.

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

The limits of the core-periphery model

Despite its great originality, the CP model has several shortcomings.

(i) The CP model explains *why* agglomeration arises but does not predict *where* this happens because the manufacturing sector may concentrate in region A or in region B when transport costs are sufficiently low. Nevertheless, however small a region advantage in size is, this region becomes the core when the snowball effect is at work. In other words, regions that were once very similar may become very dissimilar. Large and affluent regions enjoy the existence of agglomeration rents that do not easily dissipate, as illustrated by the resilience of the urban hierarchy (Eaton and Eckstein, 1997; Davis and Weinstein, 2002). However, all industries must one day decline. In this respect, the CP model does not add anything to our poor understanding of regional decline, which is not the mirror image of regional growth (Breinlich *et al.,* 2014).

(ii) The sudden and discontinuous shift from dispersion to agglomeration is an artefact of the assumption of homogeneous workers who all react in the same way to marginal variations in real wages, very much like consumers react to a marginal price undercutting in the Bertrand duopoly. Once it is recognized that individuals have different attitudes toward the non-monetary attributes associated with migration, the agglomeration process is gradual and sluggish. More importantly, workers' attachment to their region of origin act 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* (Tabuchi and Thisse, 2002).

(iii) Despite its simplicity, the welfare analysis of the CP model does not deliver an unambiguous 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 the big regions. Neither of the two configurations (agglomeration or dispersion) Pareto-dominates the other: workers living in the periphery always prefer dispersion because they do not have to import all varieties, whereas those living in the core always prefer agglomeration because all varieties are locally produced. In other words, *market integration generates both welfare gains and welfare loses through the geographical redistribution of activities.*

In order to compare the two market configurations, Charlot *et al.* (2006) use compensation mechanisms put forward in public economics to evaluate the social desirability of a move, using market prices and equilibrium wages to compute the compensations to be paid either by those who gain from the move or by those who are hurt by the move. When transport costs get sufficiently low, the winners can compensate the losers for the latter to sustain the utility level they enjoyed under dispersion. This is because firms' efficiency gains are high enough to offset the losses incurred by the 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 sharp results, even in a simple setting such as Krugman's, may explain why so many contrasting views exist in a domain where there are good reasons to believe that the underlying tenets are correct.

(iv) 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 agglomerated in a single region when the economy is multi-regional. Yet, using a discrete Fourier transformation, Akamatsu *et al.* (2012) have been able to prove the following result. Consider $K = 2^n$ regions that are equidistantly distributed along a circle. Starting from a value of the transportation costs which is large enough for the uniform distribution of the manufacturing sector to be a stable equilibrium, a gradual decrease in transport costs leads to a pattern in which workers are partially agglomerated in $K/2$ alternate regions. As transport costs steadily decrease, the CP model displays a sequence of bifurcations in which the number of manufactured regions is reduced by half and the spacing between each pair of neighboring manufactured regions doubles after each bifurcation, until the full agglomeration of the manufacturing sector into a single region. This extends the conclusions obtained by Krugman (1991), but what these conclusions become in more complex spatial settings remains an open question.

(v) Last, by neglecting that the agglomeration of activities typically materializes in the form of cities, regional economics overlooks the various costs that are typically generated by the geographical concentration of people. Yet, accounting for these costs may have a deep impact on the conclusions drawn from regional economics. For example, Helpman (1998) has argued that decreasing freight costs could well trigger the dispersion, rather than the agglomeration, of economic activities when the dispersion force is given by a given stock of housing rather than 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 price rises when workers move to the larger region, which strengthens the dispersion force. Simultaneously, cheaper transport facilitates trade. Combining these two forces shows how dispersion may arise when transport costs steadily decrease. Anticipating on what well will see in Section 4, lowering transport costs acts against the gathering of consumers within the same city/region because they allow consumers to alleviate the costs associated with the working of a city. So far, the most robust conclusion is that lowering transport costs fosters agglomeration as in Krugman and, then, leads to redispersion as in Helpman (Puga, 1999). This suffices to show that what is going on within regions is key to the understanding of the regional question.

The evolution of regional disparities: alternative approaches

Input-output linkages and the bell-shaped curve of spatial development

Moving beyond the Krugman model in search of alternative explanations appears to be warranted in order to understand the emergence of large industrial regions in economies characterized by a low spatial mobility of labor. 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 very large fraction of firms' sales, being often overshadowed by the demand for intermediates. Therefore, in making their 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 where intermediate-goods suppliers are located. 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 conversely*. Assume that many firms belonging to the final sector are concentrated in one region. The high demand for intermediate goods within this region attracts producers of intermediate goods. In turn, these intermediate goods are supplied at a lower cost in the core region, which induces even more final sector firms to move to the core. Such a cumulative causation process feeds on itself, so that the resulting agglomeration can be explained solely by the demand for intermediate goods, without having recourse to labor mobility as in Krugman's setting.

Giving intermediate goods a prominent role is a clear departure from the CP model, which allows one to focus on other forces that are at work in modern economies. To this end, note that, once workers are immobile, a higher concentration of firms within a region translates to 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 induced to relocate their activities to the periphery when lower wages there more than offset lower demand. In sum, as transport costs fall *there is, first, agglomeration and then dispersion of production*. In sum, economic integration would 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 later stages.

Technological progress in manufacturing rather than transport

The foregoing models focus exclusively on falling transport and trade costs. There is no doubt that the transport sector has faced huge productivity gains during the last two centuries. However, a great number of other sectors have also experienced spectacular productivity gains. This state of affairs has led Tabuchi *et al.* (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 because migration generates a wide range of non-pecuniary costs that have a lasting influence on individual well-being. These costs act here as the main dispersion force.

Findings differ in various respects from those obtained in the CP model. First, in Krugman (1991) and followers, the incentive to move shrinks as transport costs fall because prices and nominal wages converge. What drives Krugman's result is the change in the sign of the real wage differential when transport costs fall below some threshold. Note, however, that the absolute value of this differential steadily decreases as both the agglomeration and dispersion forces weaken with transport trade costs. By contrast, when one region is slightly bigger than the other, technological progress in manufacturing that reduces the labor marginal (resp., fixed) requirement in the two regions makes the larger region more attractive by increasing wages and decreasing the prices of existing varieties (resp., increasing wages and the number of varieties) therein. In other words, technological progress tends to exacerbate differences between the two regions and thus raises the incentive to move from the smaller to the larger region. As a consequence, technological progress in manufacturing favors agglomeration. Another major difference is worth pointing out. Falling transport costs foster dispersion here instead of agglomeration. Indeed, everything else being equal the utility differential shrinks with a deeper market integration, which incites more workers to stay put.

Last, workers move to the larger region when productivity gains are strong enough to make the utility differential greater than their mobility costs. Since these costs may vary across workers, the final pattern involves a core accommodating a higher share of firms and workers than the periphery, but this share depends on the intensity of technological progress and the level of mobility costs. In particular, high mobility costs lower the productive efficiency of the global economy but avoid increasing regional disparities. Moreover, innovations often require skilled workers who are more mobile than unskilled workers. According to Moretti (2012), “geographically, American workers are increasingly sorting along educational lines.” In this case, the size of interregional income and welfare gaps are increasingly caused by differences in the geographical distribution of skills and human capital. This has a major political consequence that should not be overlooked: *spatial inequality is associated with social polarization*. A deeper discussion of this problem would take us too far from the main purpose of this survey.

Does transportation matter for the space-economy?

According to Glaeser and Kohlhase (2004), the steady and spectacular decline of freight costs has brought us into “a world where it is essentially free to move goods.” This claim overlooks the fact that low transport costs strongly intensify competition and make firms more sensitive to tiny differences in costs. Furthermore, non-tradable services account for a large and growing share of GDPs. Hence, the importance of transport costs should be assessed with respect to the value of the tradable goods only. In addition, various transaction costs, such as time costs, that are related to distance have increased, but they are not accounted for in GDPs. Finally, by assigning different degrees of centrality to nodes in a transportation network, a specific network pattern favors some places at the expense of others. To illustrate the last point, consider the following two arguments.

First, consider the simplest location problem of a firm and ask whether a new transport infrastructure affects its location. Given a transport network defined by a set of nodes and links, 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 of the network, i.e., a market town, a resource town, or a crossroad, a result known as the exclusion property (Hurter and Martinich, 1989). This result reduces the set of all possible points along roads to the subset of nodes; it excludes all intermediate locations from further consideration. In other words, building a new transport infrastructure that connects two existing nodes will not affect the region the arc goes through, except maybe near its endpoints when the local markets are congested. On the other hand, if the new arc crosses an existing arc, the new crossroad may attract activities, and thus generate local growth and job creation.

Second, transport infrastructure has been built in West and East Africa to allow former colonies to export their mineral resources to developed countries overseas. Bonfatti and Poelhekke (2017) show that coastal countries with more mines import relatively less from neighbors than landlocked countries with more mines, because the latter need to be connected to their neighbors in order to export. This suggests that the transport networks designed during the colonial period still shape the intensity and nature of trade flows in Africa. With all this in mind, we find it hard to believe that transport networks do not matter for the organization of the space-economy. As argued for long by transport geographers (see, e.g.

Thomas, 2002), *the spatial distribution of activities depends on the shape 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 the spatial economists? There are at least two of them. First, regional economic models have the merit to show 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. That said, it is remarkable that spatial and transportation economics have developed in a rather unconnected way. The second issue deals with what seems to be an egg-and-chicken problem that generates endless debates in the popular and academic presses: does the construction of a new transport infrastructure foster regional growth, or is this infrastructure built because the corresponding regions are richer? The first question is to a large extent related to what we mean by transport costs and how we model and measure them. Since storage and transport are, to a certain extent, substitutes, what matters for firms is the level of *logistic costs*, which account for both types of costs. Freight costs have received most attention up to now but we should keep in mind that trade in services accounts for one third of world exports. Trade in services calls upon very different transport and communication channels: it can be based on electronic delivery, but it also consist of services supplied between branches of a large firm. As for the second question, it strikes us as being mainly an empirical issue that has major policy implications as the construction of a new transport infrastructure is often presented as the remedy to local backwardness.

What do we mean by transport costs?

The trade and spatial economics literature recognizes the existence of various types of spatial frictions, but assumes that an iceberg transport cost is sufficient to reflect the impact of these various frictions. The iceberg cost assumption means that freight transport costs between any two locations can be represented by a constant cost margin on top of the producer price. This assumption is made for analytical tractability: when the iceberg cost is added to the CES isoelastic demand functions, only the level, and not the elasticity of, the demand functions matters. 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 used in this literature assumes that freight costs for a given value of the load are described by the lower envelope of several linear 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 the presence of increasing returns to scale and makes shipping costs endogenous. More generally, transportation economists stress the following effects that are dismissed by the iceberg cost. First, one needs to distinguish between transport modes because they are differentiated by their market structure and technology. Second, transport rates decrease with the size of shipments (this is called *density economies*), while transport rates also decrease with the haulage distance (this is called *long-haul economies*). However, changes in transport costs are also due to exogenous forces, like strong technological advances in transport technologies and deregulation policies. Last, countries trade with themselves more than with the rest of the world, while most of the trade and new economic geography literature assumes that internal transport costs are nil. But do these simplifying assumptions matter to the spatial economists?

First of all, we want to stress that the results discussed in 3.1 and 3.2 do not depend on the use of the iceberg cost function. They hold true when additive transport costs, measured directly in the numéraire and not as a share of the good that is shipped, are combined to linear demand functions (Ottaviano and Thisse, 2004). Second, the iceberg cost assumption implies that halving the producer price of a good implies that its delivered price is also cut by 50%. In other words, the pass-through is equal to 100%. This contradicts empirical evidence and spatial price theory, which both suggest that firms adopt a freight absorption policy to ease the penetration of remote markets (Phlips, 1983). Third, working with a constant trade cost proportion amounts to assuming that the transport sector operates under perfect competition and generates a shipping rate that can be treated parametrically. This overlooks the fact that market structure matters a lot in determining freight rates. For example, Combes and Lafourcade (2005) for France and Winston (2013) for the U.S. showed that deregulation has been key in freight rate decrease.

Fourth, the iceberg cost function cannot account for density economies because the value of the cost τ_{ij} between locations 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 long-haul economies into account because the values of τ_{ij} are chosen arbitrarily. However, the use of this function in continuous location models, as in by Fujita *et al.* (1999), Rossi-Hansberg (2004) and Desmet and Rossi-Hansberg (2014), 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, while D is 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 the distance D. As a consequence, using such a function amounts to assuming that there are long-haul *dis*economies (McCann, 2005).

Last, shipping goods between two locations within the same country or between two countries involves different costs. For example, Anderson and van Wincoop (2004) show that trade costs between countries consist of 55% internal costs and 74% international costs. Since labor and capital are more mobile within than between countries, the same decrease in transport costs should be associated with different responses in firms' and workers' locations. This has led Behrens *et al.* (2007) to revisit the CP model in which each country is formed by two regions. They show that the welfare impact of trade liberalization depends on the internal geography of the two countries. In addition, density economies render the two internal geographies interdependent because they affect the intensity of trade, which in turn changes the level of transport costs.

To sum up, gathering all spatial frictions generated by trade into a single iceberg trade cost is *not* an innocuous assumption. Somewhat ironically, whereas new economic geography stresses the importance of increasing returns in manufacturing, it sets aside the fact that transportation features even stronger scale economies. Accounting for the presence of increasing returns in the transport sector should rank high on the agenda of spatial economists.

Does transport infrastructure stimulate regional economic activity?

Redding and Turner (2015) put forward in a very neat way what the main issue is: "an assessment of the economic impacts of transportation infrastructure depends fundamentally on whether changes in transportation costs change the amount of economic activity or reorganize existing economic activity."

To illustrate the nature of the difficulty, we use a simple spatial competition model in which two manufacturing firms compete in delivered prices, instead of mill prices as in subsection 2.2.3. There are two cities located at $x = 0$ and $x = 1$, as well as a continuum (or a finite number) of small places distributed in between. Firms 1 and 2 are located at $x = 0$ and $x = 1$ and produce the same good at constant marginal costs $c_1 < c_2$. Ever since Hoover (1937), it is well known that in equilibrium each firm supplies the market segment over which it has the lower delivery cost. Hence, the boundary between the two market areas is located at

$$x_m = \frac{c_2 - c_1 + t}{2t},$$

if $c_2 < c_1 + t$. Otherwise, firm 1 supplies the whole market. A Bertrand-like argument shows that firm 1 charges the delivered price $p_1^*(x) = c_2 + t(1 - x)$ at $x \in [0, x_m]$ while firm 2 sells at the delivered price $p_2^*(x) = c_1 + tx$ at $x \in [x_m, 1]$. In other words, there is spatial price discrimination.

A regional highway connecting $x = 0$ and $x = b$, with $b < 1$, is built. Over $[0, b]$ the transport rate decreases to $\bar{t} < t$. Two cases may arise. Assume, first, that $b > x_m$. Firm 1 cannot sell beyond b because its transport rate is still equal to t over $[b, 1]$. Therefore, the location of the marginal customer is on the new highway at the solution x_b to $c_1 + \bar{t}x = c_2 + t(1 - b) + \bar{t}(b - x)$:

$$x_b = \frac{c_2 - c_1 + t - b(t - \bar{t})}{2\bar{t}} < b.$$

It is readily verified that $x_b > x_m$ always holds because the highway allows firm 1 to be more aggressive on $[x_m, b]$. Under these circumstances, firm 1 benefits from the presence of a regional highway, whereas firm 2 is no longer protected by a high transport cost. However, despite the contraction of its market area, firm 2's size may grow because it sells more on any local market belonging to $[x_b, b]$. Note that all of this holds true if $c_1 > c_2$ as long as the above inequalities are satisfied. In this case, the highway is beneficial to the inefficient firm and detrimental to the efficient firm. On the other hand, when $b < x_m$, the boundary is unaffected by the highway because firm 1 benefits from a lower shipping cost over a segment that belonged to its market area before the construction of the highway. As a consequence, the highway raises firm 2's output because this firm sells at lower prices, but it has no impact of firm 1's. In sum, the effect of a regional highway on firms' output depends on the industry through its cost and demand conditions.

The reduced-form approach. In this approach, the biggest empirical challenge is probably to construct the appropriate counterfactual for the absence of the planned transport infrastructure. Chandra and Thompson (2000) conduct an analysis of the impact of interstate highways on U.S. rural counties between 1969 and 1993. They find that the net effect on regional growth is ambiguous. More specifically, a new interstate highway increases total earnings in the rural counties the highway passes through. However, it tends to cause a decline in total earnings in adjacent counties.

Consider next the effects of new infrastructure on agriculture. In two meticulous and rich papers, Donaldson (2017) and Donaldson and Hornbeck (2016) study the effect of the development of railroads in colonial India (1870-1930) and in the U.S. from 1870 to 1890, respectively. Railroads in India decreased transport costs and increased the agricultural output in the connected districts by 17%. Because railroads allowed the different regions to exploit the gains from trade, there was also an overall increase in income for India. For the US, Donaldson and Hornbeck (2016) find a strong increase in agricultural output in the districts with railroad access. Unlike manufacturing, agriculture is a dispersed activity that uses a large amount of land. It is, therefore, not totally surprising that fields located along, or close to, tracks benefited from the construction of railroads.

Let us now come to the manufacturing sector. Faber (2014) shows that the construction of new highways in China increased the industrial output of the connected metropolitan areas and decreased that of regions connected to the metros compared to not-connected regions. In other words, trade integration re-enforces the core cities at the expense of the intermediate regions. This result can be explained by adding long-haul economies to the HME model. In the same vein, Lin (2017) finds that the Chinese intercity high-speed rail has led to a significant hike in the number of cognitive jobs in connected cities, while Berger and Enflo (2017) find no conclusive evidence that the railway has fostered regional convergence in Sweden.

Unlike Redding and Turner (2015), we do not find the conclusions drawn from these papers, as well as from a few others, that conclusive. These papers differ not only in terms of estimation procedures, but they focus on different time periods (19th vs. 20th century), different countries (India vs. the U.S. or China), different transport modes operating under different technologies (railway vs. highways), and different activities (agriculture vs. manufacturing). Differences in geographical scales should also matter. India, China and the U.S. are huge compared to France or Germany. More work accounting explicitly for those differences is needed before a clear-cut answer to emerge.

Even though it is reasonable to believe the railway to be socially useful for the development of the American and Indian agriculture in the 19th century, this does not mean that building new highways will solve the today problems faced by many post-industrial regions. While one, or a few, transport infrastructure may well have a strong impact on the location and growth of activities, a large number of them is likely to have none because the quasi-ubiquity of transport infrastructure ceases to affect firms' location. In brief, though the provision of more efficient transport infrastructure may help in promoting regional growth, we believe that it does not constitute the universal panacea promoted by many policy-makers.

Reductions in travel and communication costs also affect the spatial equilibrium in various ways. The most obvious and direct effect of the decrease in passenger travel cost and time is in tourism where a large share of activities has been relocated in regions that have an absolute advantage in their natural amenity endowment (e.g. Florida or Spain). Furthermore, firms are packages of different functions, such as management, R&D, finance, and production. Due to the development of new information and communication devices, firms are now able to disperse these functions into geographically separated units in order to benefit from the attributes specific to different locations. Before the emergence of new information devices, a firm that delivered services to other regions relied on local representatives, while headquarters of multi-plant firms had local managers to whom they delegated decisions. Petersen and Rjan (2002) show how the improved information gathering on the creditworthiness of small business has increased their distance from lenders in the period 1973-1993.

The tendency of headquarters to agglomerate within large cities suggests that the steady decrease in travel and communication costs re-enforces the concentration of firms' strategic functions in large cities (Henderson and Ono, 2008). Charnoz *et al.* (2016) used the development of the high-speed railway network in France to show how the decrease in passenger travel speed between headquarters and affiliates has allowed concentrating the management functions in headquarters. For U.S. multi-plants firms, Giroux (2013) shows that the opening of a new airline connection has fostered an increase in investments in affiliates located near the airport. However, information frictions remain substantial. For example, Kalnins and Lafontaine (2013) observe that greater distance to headquarters is associated with shorter establishment longevity. Communication costs keep decreasing but they are hard to observe. Therefore, more work is called for here.

The general equilibrium approach. The big advantage of this approach 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 a new transport infrastructure can be indirectly affected through the redistribution of labor associated with the decrease in transport costs along some least-cost routes. A prominent example of the general equilibrium approach is Allen and Arkolakis (2014) who developed a continuous location model for the U.S. and use a rich and in-depth treatment of transport costs, which are modeled by taking into account all geographic details, including the least cost routes over all modes. Production is assumed to be perfectly competitive, while spatial production inhomogeneities explain why trade occurs. Allen and Arkolakis (2014) use their model to assess the effects of the interstate highway system by recomputing all bilateral transport costs without the interstate highway option. This would reduce total welfare by 1.1 to 1.4%, which means that the interstate highway system is a productive investment as a whole, even if it is far from being optimized as we will see. General equilibrium is the backbone of the quantitative spatial models which are extensively discussed by Redding and Rossi-Hansberg (2017). We will briefly discuss their main merits and pitfalls in Section 5.

The political economy of transport infrastructure. In the U.S., the federal government finances a large share of interstate highways by using revenues from the gasoline tax. Knight (2002) found that the working of the transport committees in Congress and Senate allowed building majorities for a regional allocation of funds that was highly inefficient: about half of the investment money was wasted. Redding and Turner (2015) see in the econometric evidence provided by Baum-Snow (2007) and Duranton and Turner (2012) that highways tend to be allocated to cities that grow more slowly than a randomly selected city.

In the EU, federal funds for transport investments are one of the main instruments used by the European Commission in designing its regional policy. In the late 1990s, the EU has launched a large transport infrastructure program with 30 priority projects. An ex ante assessment of this package generates three main findings (Proost *et al.*, 2014). First, only 12 of the 22 projects pass a simple cost-benefit analysis test. Second, most projects benefit only the region where the investment would 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.

When it comes to passenger transport, the EU has put a strong emphasis on high-speed rail (henceforth, HSR) investments. This contrasts with the choice made in the U.S. where air transportation for medium- to long-distance travel is used much more, while HSR projects have never taken off. De Rus and Nombela (2007) use a cost-benefit analysis to determine the level of demand that is needed to make a HSR socially beneficial. They find that a link needs some 10 million passengers per year. Many new HSRs in the EU do not meet this target. When a HSR has to cover all its costs, De Rus and Nombela (2007) find that there will be an insufficient number of passengers for the project to be economically viable. When trips are priced at marginal cost, the HSR has a better chance of passing the cost-benefit test, but charging the marginal cost requires high government subsidies. In addition, the government must be able to pick the most efficient projects, and thus cannot serve all regions equally. Note that HSR projects are also defended on environmental grounds, but sensitivity analysis shows that one needs extremely high carbon values to make HSR better than air transportation on these grounds.

The above findings illustrate the role of political economy factors in the selection of projects. The capacity of the local authorities and the quality of the planned investments are also key determinants of the success of a transport policy, but these aspects are difficult to capture in a general equilibrium model. For this we need case studies. In our opinion, using transport infrastructure as one of the main instruments to promote regional development is at best a mixed bag.

Are cities vanishing because of the Internet?

The main distinctive feature of a city is the very high density of activities and population, which allows agents to be close to one another. Households and firms seek spatial proximity because they need to interact for a variety of economic and social reasons. According to Glaeser (2011), the main reason for the existence of cities is to *connect* people. 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 to 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 highlights the following fundamental principle: the population distribution is the outcome of the trade-off between the human propensity to interact with others through a variety of mechanisms - the centripetal force - and different congestion effects - the centrifugal force.

In accordance with the fundamental trade-off of spatial economics, urban economics emphasizes the arbitrage between increasing returns external to firms and people's commuting, while regional economics focuses on internal increasing returns and the shipment of commodities between regions. It is, therefore, no surprise that, to a certain extent, results in regional and urban economics bear some resemblance.

Agglomeration economies

It is well known that households in large metropolises pay high rents, have a longer commute, live in a polluted environment, and face high crime rates. So why do they choose to live in such places? It is because they get much better pay in large cities than in small towns. But why do firms pay higher wages to their employees? 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 because a great number of advantages are associated with a high density of activities. These advantages are gathered under the nickname 'agglomeration economies'.[4] They involve both pecuniary and non-pecuniary external effects while they can be intrasectoral or intersectoral. For a long time, agglomeration economies were used as black boxes hiding rich microeconomic mechanisms that lead to increasing returns at the aggregate level. Nowadays, these boxes have been opened and we have a much better understanding of these various mechanisms, though their relative importance remains an unsolved empirical question. Notwithstanding the immense interest of those contributions, it is worth stressing that urban economists pay little attention (if any) to the market structure of the industry they study, although it may vary a lot across sectors. For example, the typical case of 'co-opetition' involves firms sharing knowledge and competing fiercely on the product market in a repeated game context.

The nature of business agglomeration economies

Agglomeration economies appear under very different disguises. 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 a local production function displaying increasing returns.

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

(ii) Matching means that the number of opportunities to better match workers and job requirements, or 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. Because they face a large number of potential employers, workers living in large cities do not have to change place to switch to another employer. This makes these workers more prone to change jobs. Therefore, workers having the same skills will earn higher wages in larger than in smaller cities because firms have less monopsony power (Manning, 2010). Hence, a larger labor market also makes workers less prone to change occupations (Bleakley and Lin, 2012).

(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 own 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 us, however, that research and innovation are among the most

[4] The idea of agglomeration economies dates back to Marshall (1890) and Hoover (1936). 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.

geographically concentrated activities in the world (Feldman and Kogler, 2010). This is only a seeming 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 (Gaspar and Glaeser, 1998; Glaeser, 1999; Leamer and Storper, 2001).

Research and development often demands long periods of exchange and discussion, during which knowledge is gradually structured through repeated trial and error. Thus, extensive and repeated informal contacts between agents located close to one another facilitate the diffusion of new ideas and raise the level of coordination and trust. It then becomes possible to guess why innovation is geographically concentrated: greater creativity is possible when researchers gather together. Different estimation procedures of the spatial extent of knowledge and information spillovers, such as those by Arzaghi and Henderson (2008), Belenzon and Schankerman (2013), and Buzard *et al.* (2015), suggest that *physical proximity still matters for the diffusion of information*, probably because learning from other people is the easiest way to know what is going on, but through channels that may vastly differ according to the activity under consideration. As nicely summarized by Glaeser (2011), even in the era of the Internet, "ideas cross corridors and streets more easily than continents and seas."

Equally important, 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, while Bacolod *et al.* (2009) and Combes *et al.* (2015) observe that the urban wage premium associated with large cities stems from cognitive skills rather than motor skills. Thus, everything seems to work as if the marginal productivity of a worker endowed with a certain type of skill would increase with the number of skilled workers working or living around. Therefore, it is no surprise that high-skilled workers tend to sort out according to bigger and more expensive cities. This evolution would have the following implication: cities specialized in high-tech industries attract high-skilled workers, who in turn help make these places more successful. In this context, *spatial inequality reflects more and more differences in the distribution of skills and human capital across space.*[5].

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 would still be the best locations for information-consuming activities, especially when firms operate in an environment of rapid technological change and fierce competition. Nevertheless, the importance of face-to-face contacts is likely to remain a highly debatable issue as such contacts are difficult to observe while it is even harder to assess their contribution to the process of innovation.

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 (Moretti, 2012). Thus, even if the spatial

[5]See Glaeser and Maré (2001), Moretti (2012) and Diamond (2016) for the U.S., Combes *et al.* (2008) for France, Mion and Naticchioni (2009) for Italy, and Gibbons *et al.* (2014) for the U.K.

concentration of human capital boosts economic and technological development, it might also come with a strong *regional divide*, which is likely to have political and social consequences that should not be underestimated.

How to measure business agglomeration economies?

Starting with the highly influential work of Glaeser *et al.* (1992), Henderson *et al.* (1995), and Ciccone and Hall (1996), research on city size, employment density, and productivity has progressed enormously during the last two decades. An ordinary least squares (OLS) regression of the logarithm of the average wage on the logarithm of the employment density across cities yields an elasticity that varies from 0.03 to 0.09 (Rosenthal and Strange, 2004). Hence, doubling the employment density would be associated with a productivity increase varying from 2 à 6.5%. However, there are good reasons why these results should be approached with extreme caution because some econometric problems have not been properly addressed (Combes and Gobillon, 2015). Most importantly, the estimation of the impact of local variables is inevitably plagued by reverse causality because high-wage places attract workers, thus making the employment density endogenous.

First, using a simple reduced form omits explanatory variables whose effects could be captured by the employment density. For example, overlooking variables that account for differences in, say, average skills or 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. Rather, using city and industry fixed effects, as well as individual fixed effects when individual panel data are available, 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 also 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, on their density of employment. As a consequence, employment density is correlated with the dependent variable and, therefore, with 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 to correct endogeneity biases, whether they result from omitted variables or reverse causality, involves using instrumental variables. This consists of finding variables that are correlated with the endogenous explanatory variables but not with the residuals.

Taking into account additional explanations of workers' productivity (such as non-observable 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 less than 0.03. This elasticity measures the *static* gains generated by a higher employment density (Combes *et al.*, 2012). Moreover, De la Roca and Puga (2017) also highlight the existence of *dynamic* gains stemming from the accumulation of experience individuals build when they work in large cities, which come on top of static gains.

Even though it is not disputable that agglomeration economies do exist, several issues remain unclear as it is hard to test predictions that are unique to a specific agglomeration economy, such as those discussed above, and not associated with another. In a recent comprehensive study, Faggio *et al.* (2017) give a qualified answer to these questions. They confirm the presence of the various effects discussed above but stress the fact that agglomeration is a very *heterogeneous* phenomenon. For example, low-tech industries benefit from spillovers, though less than high-tech industries. Both intrasectoral and intersectoral 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 would benefit more from spatial proximity than larger firms.

Despite the wealth of new and valuable results, it is fair to say that the dust is not settled yet. If we want to design more effective policies for city development or redevelopment we need a deeper understanding of the drivers that stand behind the process of agglomeration in cities that vastly differ in size and in their historic and geographic attributes. For example, does it make sense to expect New York City and Des Moines (Iowa) to have the same portfolio of agglomeration economies? Probably not. 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 (see, e.g. Ahlfeldt *et al.*, 2014).

Consumer agglomeration economies

The usual cliché is that big cities are bad for consumers. But the authors of anti-city pamphlets forget two things: (ii) all over the world, free people vote with their feet by moving to cities; and (ii) cities are also great consumption, culture, and leisure places (Glaeser *et al.*, 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 as the number of shops, cultural amenities, and opportunities for social relations all increase with city size.

While the Industrial Revolution fostered the emergence of manufacturing cities, services continue to show a taste for cities that manufacturing sectors no longer have. The access to a wide diversity of tradable and non-tradable goods and services is a major asset to consumers who have a preference for variety and/or display heterogeneous tastes. Even in the absence of trade, consumers and firms may choose to be agglomerated when the range of non-tradable services is sufficiently wide. Moreover, as the steady decline in transport costs has vastly improved the access to foreign goods, the resulting increase in competition has incentivized firms to restore their profit margins by supplying higher-quality goods as well as a wider range of varieties, which are available when the population is large and heterogeneous (Berry and Waldfogel, 2010; Schiff, 2015).

Tradable goods may also be less expensive in large than in small cities. Since a larger city provides a larger outlet for such consumption goods, there is more entry, which intensifies competition. Moreover, large cities attract the most efficient retailers that also benefit from better logistics. As a result, everything else being equal, market prices of tradable goods would be lower in larger cities than in smaller cities. Calculating an urban price index for 49 U.S. cities, Handbury and Weinstein (2015) find that prices fall by 1.1% when population doubles, while the number of available products increases by 20%. However, when their productivity is not positively affected by the density of activities, workers producing non-tradable services must be paid a higher wage to compensate them for the higher housing and commuting costs they bear in a bigger city. Therefore, the price of non-tradable services is likely to be higher. But this is not yet the end of the story, for the quality of those services may be higher in larger cities. The total impact of city size on the cost of living is thus a priori undetermined. More work is called for here.

Last, a large number of people facilitate the provision of public goods that could hardly be obtained in isolation because these goods would be supplied at a level inferior to the critical mass that permits them to deliver their full impact.

The trade-off between commuting and housing costs

The monocentric city model

Cities also have a bad side that puts a cap on their size. Indeed, the positive effects associated with city size come with different negative effects, such as expensive housing, long commutes and traffic congestion, pollution, and crime. Cities may therefore be viewed as the outcome of *the trade-off between agglomeration economies and various spatial frictions*.

The first analysis of the way land is allocated across different activities was by von Thünen (1826) who is considered as the founding father of spatial economics. The authoritative model of urban economics, which builds on von Thünen, is the featureless monocentric city model in which a single and exogenously given central business district (henceforth CBD) accommodates all jobs. In this context, the only spatial characteristic of a location is its distance from the CBD. The main purpose of this model is to study households' trade-off between housing size - which is approximated by the amount of land used - and their accessibility to the CBD - which is measured by the inverse of the commuting costs.

Since they dislike long trips, consumers compete for land with the aim of being as close as possible to the CBD. However, since they also prefer big lot size, consumers will not get packed in the CBD vicinity. Some of them will commute over long distances. Although consumers compete for land, space is endowed with an exogenous inhomogeneity - the CBD - which renders a consumer's equilibrium locations independent of the others'. This is why a perfectly competitive land market may sustain a distribution of consumers across locations in which no consumer can be strictly better off in another location. The argument is disarmingly simple.[6]

Consider a one-dimensional space $[0, \infty)$ 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 forces outside the city. Denoting by $R(x)$ the land rent prevailing at x and by $T(x)$ the commuting cost borne by a consumer residing at x, the budget constraint is given by $z(x) + h(x)R(x) = I(x) \equiv Y - T(x)$.

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

$$h(x)\frac{\mathrm{d}R}{\mathrm{d}x} + \frac{\mathrm{d}T}{\mathrm{d}x} = 0. \tag{2}$$

Since a longer commute generates a higher cost ($\mathrm{d}T/\mathrm{d}x > 0$), this condition holds if and only if the land rent decreases with the distance to the CBD. As a consequence, (2) means that a marginal increase in commuting costs associated with a longer trip is exactly compensated by the marginal drop in housing expenditure. To put it bluntly, people trade bigger plots for higher commuting costs. If commuting costs were independent of the distance ($\mathrm{d}T/\mathrm{d}x = 0$), the land rent would be constant and equal to 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 land rent is positive is the physical proximity to the CBD.

Furthermore, the lot size occupied by a consumer must increase with the distance from the CBD. Indeed, although a longer commute is associated with a lower net income $Y - T(x)$, the consumer optimization problem yields a compensated demand for land that depends on the land rent and the endogenous utility level which 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

[6]Note that Alonso (1964) and Fujita (1989) have developed the monocentric city model by building on von Thünen's idea that land users (farmers in his original setting) behave as if they were involved in a gigantic auction. More specifically, given her income and preferences, a consumer is characterized by a *bid rent function* that specifies her 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 (there exists no second-highest bid) and, in consequence, the land rent is the upper envelope of consumers' bid rent functions.

words, as the distance to the CBD increases, the lot size rises whereas the consumption of the composite good decreases. This, in turn, implies that the population density decreases with the distance from the CBD.

In spite of its extreme simplicity, the monocentric city model tells us something important: when more consumers get 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 cost and the opportunity cost of land occupied by the urbanites, which are all needed for consumers to reach the utility level u:

$$C(N;u) = \int_0^B [T(x) + Z(h(x);u) + R_0]\, n(x)\mathrm{d}x \qquad (3)$$

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). As a consequence, for the utility level to remain the same, it must be that *the average urban cost borne by the incumbents increases with the arrival of new residents.* In other words, regardless of the shape of $T(x)$, commuting costs generate *agglomeration diseconomies* with respect to population size. Increasing the opportunity cost of land leads to more concentrated populations and less well-being for consumers. Under such circumstances, the land rent level reflects not only the proximity to the CBD, but also the "artificial scarcity" of land stemming from restrictive land use regulation, the provision of open spaces, or public policies that maintain the prices of agricultural produces far above the international level.

In the same spirit, the implementation of urban containment hurts new residents by reducing their welfare level or motivates a fraction of the city population to migrate away (Glaeser *et al.*, 2006). In addition, by restricting 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. For example, Cheshire *et al.* (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. Even more surprising, Hsieh and Moretti (2015) find that lowering constraints on housing supply in very productive cities such as New York, San Francisco and San Jose would increase the American GDP by 9.4%, which is astronomical. It is clear that more work is called for to assess accurately the social cost of the gallery of land and housing regulations implemented in many countries. Nevertheless, we may be confident that this cost will be everything but small.

Contrary to a belief shared by the media and the public, the rise in housing costs in many cities is driven mainly by an excessive regulation of the housing and land markets.

Public policies typically place a strong constraint on the land available for housing and offices. By instituting 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. The beneficiaries of these restrictions are the owners of existing plots and buildings. Since the marginal urban cost $\mathrm{d}C(N;u)/\mathrm{d}N$ grows, young people and new inhabitants, particularly the poorest, are the victims of these price increases and crowding-out effects, which often make their living conditions difficult.

The monocentric city model has produced a great variety of results that are consistent with several of the main features of cities. The best synthesis of what has been accomplished remains the landmark book of Fujita (1989). The basic model of urban economics disregards several key aspects of a city, including heterogeneous consumers and a space differentiated by natural and historical amenities (Brueckner *et al.*, 1999). More importantly, it remains silent on why jobs are geographically concentrated. Indeed, the existence of several employment centers within the city allows alleviating urban costs. To illustrate, consider the example of the fixed lot size, whence the consumption of the composite good is the same across locations ($h = 1$), while commuting costs are linear in distance ($T(x) = tx$). When the city is monocentric, the level of urban cost is obtained by integrating $R(x) + tx = tN$ over $[0, N]$, that is,

$$C_1(N, u) = tN^2,$$

while the urban cost becomes

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

when the city has two employment centers located at $x = 0$ and $x = N$, respectively. Since $C_2(N, u) < C_1(N, u)$, the equilibrium utility level would be 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 are the usual suspects.

The emergence of employment centers

The first answer to the above question was provided by Ogawa and Fujita (1980) in a fundamental paper that went unnoticed for a long time, probably because the field was still at the periphery of the economics profession. These authors use a gravity-like reduced form for spatial externalities and combine consumers and firms in a full-fledged general equilibrium model in which goods, labor, and land markets are perfectly competitive. Spatial externalities act as an agglomeration force because their intensity is subject to distance-decay effects. However, the clustering of firms increases the average commuting distance for workers, which in turn leads workers to pay a higher land rent. Therefore, firms must pay workers a higher wage as compensation for their longer commutes. 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 those opposing forces. Note the difference with the monocentric model in which the CBD is given: interactions among agents make the relative advantage of a given location for an agent dependent on the locations chosen by the other agents.

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

$$Y(x) = Y - \int_{-b}^{b} \tau|x-y|m(y)\mathrm{d}y, \tag{4}$$

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, like in pre-industrial cities when people moved on foot, the equilibrium involves a complete mix of business and residential activities with everyone living where he or she works. In this case, land is unspecialized. As commuting costs fall, two employment centers, which are themselves flanked by a residential area, are formed around a district in which firms and workers are uniformly mixed. Eventually, when the emergence of mass transport means and the use of car commuting costs get low enough, the interior district vanishes and the city becomes monocentric. In this case, land is fully specialized between residential or commercial activities. We summarize as follows:

The city structure. *Assume a linear spatial externality 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 $\tau K < t$.*

Evidently, 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 spatial externalities 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 CBD center.

Given the importance of the subject matter, it is surprising that only a handful of papers have explored more general or alternative settings. This makes it hard to adopt a structural approach in the study of knowledge spillovers (Combes *et al.*, 2012). Fujita and Ogawa (1982) consider a negative exponential decay-function and show by using simulations that polycentric configurations exist. However, there is multiplicity of equilibria. Lucas and Rossi-Hansberg (2002) also use an exponential decay-function but consider a neoclassical production function involving land and labor. Roughly speaking, their simulations yield results consistent with those obtained by Fujita and Ogawa (1982). Since firms operate under constant returns, Berliant *et al.* (2002) replace the firm density by the aggregate capital stock, denoted $\bar{K}$. More precisely, a firm produces under a Cobb-Douglas production function:

$$Y(x) = K^{\alpha}L^{\beta}[a(x)\bar{K}]^{1-\alpha-\beta},$$

the spillover effect being now given by $a(x) = \alpha - x^2 - \beta\sigma^2$ where σ is the absolute deviation of the firm distribution. Berliant *et al.* (2002) show that the equilibrium displays one of the three configuration identified by Ogawa and Fujita (1980). Thus, the functional form of the spatial externality function matters for the city structure. Assuming decreasing and convex functions is not sufficient to determine the nature of the equilibrium. Although the negative exponential seems to be a natural candidate, the sensitivity of results to the specification of the spatial externality is evidence that more work is called for.

Congestion costs

Complaining about transport conditions is as common as chatting about the weather. The origin of the discomfort lies in the various negative external costs experienced in most urban trips. The main external travel costs within cities are due to road congestion, followed by local air pollution, accidents, and climate effects. That neither the U.S. nor the EU have 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 considered as 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, we find it important to distinguish between two fundamentally different instruments: on the one hand, a better use of existing transport infrastructure through pricing and regulation and, on the other, the addition of transport capacity.

Reducing congestion via road pricing

Ever since the pioneering work of Pigou (1920), there has been a general (but not universal) agreement among economists that road pricing is the ideal instrument to tackle urban road congestion. The argument is straightforward. Beyond a certain traffic density, travel speed falls as the number of car-drivers increases. Although the costs of travel delay are borne by drivers collectively, each individual driver neglects the external cost of delay they impose on other drivers. 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 φ. A population of N homogeneous users residing in A wishes to travel to B at the same time, but the capacity of the road is insufficient to allow this. In the simplest formulation, the (average) cost of a trip is given by the function $ATC(N) = \alpha N/s$, where s is the capacity of the road and the cost φ has been normalized to zero. 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. Moreover, travelers do not have to depart at the same time. In the bottleneck model developed by Vickrey (1969) and subsequent authors, the road has a bottleneck with a flow capacity of s cars per unit time. If the rate at which vehicles arrive at the head of the bottleneck exceeds s, a queue develops. Drivers have a common preferred or ideal time to arrive at their destination, and incur a so-called *schedule delay cost* if they arrive earlier or later. The cost of a trip (with $\varphi = 0$) is the sum of the queuing delay cost, the schedule delay cost, and the toll (if any). Let α denote the unit cost of queuing time, β the unit cost of early arrival (with $\beta < \alpha$), and γ the unit cost of late arrival. Let $\tau(T)$ denote the toll levied at time T, and $Q(T)$ the number of vehicles in the queue at time T. A driver who departs at time T incurs a trip cost of $C(T) = \alpha Q(T)/s + \beta(\text{time early}) + \gamma(\text{time late}) + \tau(T)$.

Since drivers are homogeneous, the equilibrium travel cost must be the same throughout the period during which drivers depart. If there is no toll, 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 *et al.* (1993) show that the equilibrium travel cost is equal to $\delta N/s$ where $\delta \equiv \beta\gamma/(\beta+\gamma)$. The equilibrium is inefficient because queuing time is a deadweight loss. Queuing can be prevented by levying a continuously time-varying or *fine* 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. Queuing cost is eliminated, while schedule delay costs are unchanged because 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 fine toll the total social variable costs of travel are halved without reducing the number of trips made at all. This in turn implies that the benefits associated with density are less affected (Arnott, 2007). However, the costs of congestion are not eliminated because schedule delay costs are unchanged. These costs are hidden in Pigou's static model.

So far, we have assumed that car users have the same value of time. However, the empirical evidence suggests that individuals are very heterogeneous in their values of travel time (Small *et al.,* 2005). In this case, by substituting high value trips (business and highly-skilled commuters) for low value trips (leisure), road pricing generates an additional benefit. Indeed, by reducing the number of drivers during the peak period, road pricing favors productivity by making business-to-business trips cheaper. Moreover, there is a growing evidence that individuals who have a long commute are more prone to being absent from work, to arrive late at the workplace and/or to make less work effort (Van Ommeren and Gutiérrez-i-Puigarnau, 2011).

The difficult road to congestion pricing

Congestion pricing has been studied intensively in transportation economics. However, if we have a fairly good understanding of the main issues at stake when agents have fixed locations, the literature does not have much to say about the locational effects of congestion pricing. Three main lessons can be drawn from the state of the art. First, the design of the road pricing scheme is very important for the magnitude of the total net welfare effects (Anas and Lindsey, 2012). For example, Stockholm has implemented a more efficient scheme than London: the system has lower transaction costs and uses more finely differentiated charges over the time of the day. Indeed, as shown by the bottleneck model, time differentiation is crucial for 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 trips to alleviate congestion. Second, in the 10 to 20% reduction in car use necessary to eliminate most queues, only a relatively small share (40% or less) of the suppressed car trips is replaced by mass transit; the rest of the trips disappeared due to car sharing, combining trips, or simply foregoing the trip (Eliasson *et al.*, 2009).

Last, standard cost-benefit analyses (henceforth, CBA) are confined to effects within the transport sector. However, urban economics suggests that there are *wider benefits* caused by a better accessibility, which are in line with those associated with a higher density. For example, Anderstig *et al.* (2016) estimate a reduced-form relationship between accessibility and labor income for Stockholm. High-income earners are found to benefit more from an increase in accessibility via higher income gains, while some low-income groups experience a small drop in income. These effects are congruent with the agglomeration benefits discussed above, except that we now assess the effect of accessibility rather than density. When the accessibility gains generated by the Stockholm cordon toll in 2008 are simulated on the same dataset, Anderstig *et al.* (2016) find that total income gains could well be as important as the direct time benefits. Thus, a standard CBA of congestion pricing focusing on time benefits only could vastly under-estimate the total benefits of substantial improvements in accessibility.

Recent papers shed a contrasting light on the merits of congestion pricing when it is recognized that agents can change location in response to substantial changes in travel costs. For example, Hymel (2009) considers the 85 largest U.S. metropolitan areas and finds a negative correlation between employment growth and congestion measured by the average number of hours lost in traffic. Simulating a reduced-form model, Hymel finds that a 50% reduction in traffic congestion would increase employment from 4 to 11%. Brinkman (2016) estimates a spatial equilibrium model that includes congestion costs and agglomeration economies. In the calibration of his model to Columbus (OH), congestion pricing may have negative effects because the congestion tax, although effective in reducing congestion, weakens the agglomeration effects through a more dispersed employment. It seems premature to draw firm conclusions from a handful of papers that use the conventional congestion model with homogeneous drivers. Clearly, congestion pricing has various effects that need to be studied more carefully.

In the U.S., where road pricing seems to be banned from the public debate, there is more focus on optional varieties of road pricing like pay lanes. As pay lanes send motorists to an unpriced alternative (the other lanes), pay lanes can only generate net welfare benefits when either the car users differ in their value of time or when there is a fine-tuned pricing of the bottleneck (Small and Yan, 2001). In the EU, only a few cities (London, Stockholm, Milan, Göteborg) have implemented congestion pricing schemes. Most national and local governments alike favor other policies such as high gasoline prices, as well as large investments and subsidies in mass transit. So, one may wonder why, despite high potential benefits, road pricing is so unpopular?

De Borger and Proost (2012) proposed a political economy analysis. The population is a priori divided into three categories: the non-drivers; the drivers who can easily switch to transit (who are called marginal drivers); and those who face high switching costs. When toll revenues are evenly redistributed across people, non-drivers support congestion pricing. Therefore, non-drivers and marginal drivers could form a majority for congestion pricing. However, if all drivers know only the average switching costs to public transports, the marginal drivers expect to bear a cost that might be much higher than what it 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, the marginal car users know that their switching costs are lower than what they expected, and thus may support congestion pricing ex post. Hence, a majority of drivers may vote against road pricing ex ante and even against an experiment because they view their expected gains as negative, whereas a majority may support it when implemented. As shown by the examples of London and Stockholm, the final decision depends on the local government's ability to organize an experiment.

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 transport 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 $SMC_{PT} - P_{PT} > 0$ is able to make car users switch to public transport. More specifically, for a given subsidy, the fraction φ of new transit passengers who would, in the absence of the subsidy, be car users, must satisfy the following relationship at peak time:

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

Parry and Small (2009) have found that a subsidy close to 90% 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 to 10%.

Rail and metro systems display strong increasing returns to scale. 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 that characterizes almost all public transport systems. Bus systems exhibit smaller scale economies, so that an accurate peak load pricing can make them break-even more easily. It should be clear that there is a need for more efficient pricing systems. These ones should account for the differences in cost between peak and off-peak trips and vary with 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.

Does 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 the road capacity is expanded: the new capacity attracts more car users. Eventually, expanding road capacity may create its own demand, a phenomenon known as the Pigou-Knight-Downs paradox (Arnott and Small, 1994). This paradox is nothing else than a demand for transportation that is elastic with respect to the level of travel costs. But is this paradox more than an intellectual curiosity? Duranton and Turner (2011) have revisited the problem in the case of American cities for the years 1983, 1993, and 2003, while paying a 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." To put it differently, the elasticity of road use with respect to the number of urban lane-kilometers is close to one, thus making the Pigou-Knight-Downs paradox a result. But 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, the road extension attracts mass transit passengers. This reduces frequency in public transport, which in turn increases waiting and schedule delay costs, a vicious circle that may lead to the disappearance of the transit alternative. Eventually, roads will attract even more users.

Very much like decreasing transport costs affect firms' locations, large projects that substantially lower commuting costs are likely to affect households' residential choices. For example, Baum-Snow (2007) found that between 1950 and 1990, a new highway passing through a central city reduces its population by about 18%. His estimates imply that the aggregate central city population would have grown by about 8% had the interstate highway system not been built. While central cities were still the origin or destination of 66% of all commutes in 1960, in 2000 this share has dropped to 38%. This suggests that jobs have followed the residents in their suburbanization. However, care is needed because several effects are at work here (Brueckner, 2000). In the same vein, Garcia *et al.* (2015) looked into the effects of highways on urbanization patterns in Spain. They found that a highway emanating from central cities caused an $8-9\%$ decline in central city population between 1960 and 2011. In addition, a highway ray fostered a 20% population growth in the suburban municipalities where ramps were located. Last, each additional kilometer close to the nearest highway ramp increased municipal density growth by 8%. All of this confirms the impact of increasing highway capacity on the population distribution within metropolitan areas.

The foregoing results have two major implications that run against many policy recommendations: when road pricing is not implemented, building new roads need not be the appropriate policy to reduce traffic congestion. By contrast, 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 those generated by new transportation projects.

Economists have developed CBA techniques that aim to assess the desirability of transport projects. The CBA techniques have progressed over the last 50 years from the Dupuit consumer surplus to methods that correct for externalities, as well as for 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). More specifically, there is a need for operational models integrating both land use and transport (LUTI). Given the long run implications of decisions made about land use and transport infrastructure, the market alone cannot solve all problems. Accordingly, *cities need to be planned*. For this, different agents (developers, firms, governmental agencies) pursuing different, and sometimes conflicting, objectives must coordinate their actions. Furthermore, coordination requires commitment on the part of some agents, which is not always possible.

Therefore, developing LUTI models is a formidable challenge. It is only recently that researchers have tried to build such models in line with the basic principles of urban economics and general equilibrium theory (Anas and Liu, 2007).

To summarize, we begin to understand the different mechanisms that come into play: agglomeration economies, congestion, environmental externalities, as well as the impacts of policy instruments (land use, buildings regulation, transport and parking pricing, and road capacity). However, our knowledge is still very partial, as most studies focus on only one or two mechanisms and only one instrument at a time. However, we may already safely conclude that 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 and expensive transportation infrastructures, and dampen the sprawling of activities.

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 to figure out how forces acting on different spatial scales shape the global economy. Most countries trade more with themselves than with the rest of the world while cities are major actors in the process of trade. It is, therefore, fundamental to understand how the intensity of trade is influenced by the size and structure of cities and, conversely, how trade and market integration affect the internal structure of cities. This task must be accomplished within a multi-city framework to capture as many general equilibrium effects as possible. In doing so, however, we cannot ignore the relative position of cities, as expressed by different accessibility measures such as transport costs.

Not all cities are alike: why is the urban system hierarchical?

The most enduring problem, that is, the existence of an *urban system* involving large and medium-sized cities as well as towns and villages, which produce and trade different commodities remains largely unsolved. Henderson (1974, 1988) has developed a compelling and original approach that allows one to describe how an urban system involving an endogenous number of specialized cities of different sizes which trade commodities. In each city, there is again a tension between two forces. On the one hand, there are external economies associated with the agglomeration of firms at the CBD. On the other hand, there are diseconomies generated by the need to consume land and to commute to the CBD. The market for cities is characterized by competition among *land developers* or *local governments*, which understand that they may benefit from organizing cities in a way that maximizes the utility level of residents, while internalizing the external effects generated by the agglomeration of firms and workers belonging to the same industry. In equilibrium, the utility level is the same across cities and each city has a positive and finite size. As cities vary in their industrial specialization, they have different sizes because industries differ in the external economies they are able to create.

To shed light on the forces at work in the urban system, we build on Fujita and Thisse (2013) who solve Henderson's model as a full-fledged general equilibrium model under the following assumptions. Consider a mass $\mathbf{N}$ of identical consumers endowed with one unit of labor and Cobb-Douglas preferences $U = x_1^{\alpha_1} \cdots x_n^{\alpha_n} s$, where $s = 1$ stands for the fixed lot size. The production function of sector $i = 1, ..., n-1$ is given by $F_i(N) = N^{1+\gamma_i}$ where $0 < \gamma_i < 1$ is the degree of increasing returns in sector i, while n is produced under constant returns and is chosen as the numéraire. Goods are indexed for $\gamma_1 > ... > \gamma_n = 0$ to hold. Commuting costs are linear in distance, and thus tx units of the numéraire are needed to cover the distance x. Last, trading commodities between cities is costless. This assumption implies that each commodity is available at the same price regardless of the city in which consumers live. Since consumers enjoy the same utility level, they have the same consumption structure.

Fujita and Thisse show that the equilibrium involves

$$m_i^* = \alpha_i t \frac{(1-\gamma_i)^{1+\gamma_i}}{\gamma_i} \frac{\mathbf{N}}{4} \qquad i = 1, ..., n-1 \tag{6}$$

type-i cities whose size is given by

$$N_i^* = \frac{4}{t} \frac{\gamma_i}{1-\gamma_i}, \tag{7}$$

while n is produced in a very large number of arbitrarily small cities. Thus, increasing returns prevent the proliferation of cities ($m_i^* \to \infty$ when $\gamma_i \to 0$), whereas commuting costs forestall cities to be indefinitely large ($N_i^* \to \infty$ when $t \to 0$). Since the degree of increasing returns varies across goods and services, *cities specialized in the production of different commodities have different sizes.* In particular, (7) implies that large (small) cities are those specialized in the production of commodities with high (low) degrees of increasing returns: $N_{n-1}^* < ... < N_1^*$. However, the number of large cities need not be smaller than the number of small cities because α_i may be larger than α_{i-1}.

It follows from (6) that the number of type-i cities decreases with the degree of increasing returns γ_i because a smaller number of larger cities allow for a better exploitation of scale economies. Low commuting costs permit the coexistence of cities having different sizes, in which workers are paid different wages but pay different land prices. By contrast, when commuting costs were very high, urban systems are predominantly formed by small cities that do not differ much, like in the pre-industrial times. Once more, we see that the fundamental trade-off between agglomeration economies and commuting costs shapes the urban system: (i) as increasing returns get stronger in the production of commodity i, the number of type i-cities decreases, but these cities become larger; and (ii) when commuting costs fall, all cities get bigger, while the number of each type of cities decreases.

The actual number and size of cities seem to obey a simple, but intriguing, empirical rule, which keeps attracting 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 territory. To put it differently, if there is a single type 1-city, the two type 2-cities have one-half the population P_1 of the largest city, the three type 3-cities host one-third of that population, and so on. Formally, the rank-size rule holds that

$$\ln r_i = \ln P_1 - b \ln P_i$$

where P_i is the population of cities of rank i in the urban hierarchy and $b = 1$. A large number of estimations of b suggest a value close to 1. It is now recognized that a power function provides a good approximation of the population distribution (Gabaix, 2009). But is $b = 1$ the best estimation? The pooled estimate of this coefficient 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 solution of Henderson's model sheds light on the rank-size rule. Multiplying m_i^* and N_i^* yields the total population P_i^* living in type-i cities. Renumbering the sectors for $P_1^* > ... > P_{n-1}^*$ to hold, we obtain:

$$\ln P_1^* - \ln P_i^* = \ln \alpha_1 + \gamma_1 \ln(1 - \gamma_1) - \ln \alpha_i - \gamma_i \ln(1 - \gamma_i). \tag{8}$$

Since this expression involves demand and supply side parameters that are a priori independent, there is no reason to expect (8) to be (more or less) related to the log of the integer i. Hence, we find it fair to say that the rank-size rule is a mystery that is likely to remain so.

The second generation models of urban system mainly focus on the sorting of heterogeneous workers across cities. They aim to provide a description of the relationship between the size of cities and their skill compositions which is consistent with empirical evidence (Behrens *et al.*, 2014; Eeckhout *et al.*, 2014; Davis and Dingel, 2017). Importantly, unlike Henderson, these models show how different cities may emerge without having to appeal to 'large agents' such as developers or local governments. Davis and Dingel (2017) are the first who show how costly face-to-face contacts across heterogeneous workers, the importance of which has been stressed above, may drive the emergence of an urban hierarchy. More specifically, these authors consider a noncooperative game in which workers endowed with different abilities allocate their time between producing tradables and exchanging ideas. Key to their analysis is that individual productivity depends on personal skill, but also on the quality of the exchange-idea environment which is determined by the composition of the population involved in the local exchange process. Furthermore, individual skills and the quality of the idea-exchange environments are complements. Under these reasonable assumptions, Davis and Dingel (2017) show that there is spatial sorting along the skill dimension, where large cities are more skill-abundant while skill premia are higher in bigger cities. In the spirit of the CP model, stable equilibria have equal-sized cities when the agglomeration force is weak relative to urban costs.

Unfortunately, those models keep assuming that cities produce the same commodity that can be traded costlessly between cities. Like in Henderson, they do not recognize that cities are anchored in specific locations and embedded in intricate networks of trade relations that partially explain their size and industrial mix. In other words, cities are like "floating islands". A comprehensive theory of urban systems should explicitly account for city location. For this, we must account for positive transport costs between cities. However, doing this turns out to be a hard task because this implies that commodities are sold at different prices in different cities. In this case, solving the model analytically becomes problematic. One of the few attempts we are aware of is the work of Murata and Thisse (2005) who developed a setup in which each region is described by a monocentric city. Local and global forces then interact to determine the geography of production and employment as the size of the two cities depends on the interplay between commuting costs *within* cities and transport costs *between* cities. Cavailhès *et al.* (2007) go one step further by recognizing that firms may alleviate the burden of urban costs in large metropolitan areas through the emergence of secondary employment centers (Henderson and Mitra, 1996). In this way, firms are able to pay lower wages and land rents while retaining most of the benefits generated by large urban agglomerations. However, for this to happen, firms established in the secondary centers must maintain a very good access to the inner city where specialized business-to-business services are supplied. As a consequence, another type of spatial friction comes into play, that is, *communication costs*. In this context, agglomeration and dispersion take two distinct forms because they are now compounded by the centralization or decentralization of activities within the same city. Therefore, the development of new information and communication technologies may prevent the re-dispersion of activities between cities that a deep economic integration is expected to trigger through the bell-shaped curve of spatial development. The merit of such models is to bring to the fore new effects that stem from the blending of regional and urban economics. Their main shortcoming is that they remain confined to two-city settings.

Quantitative spatial economics

The lack of an analytical solution to the dimensionality problem discussed above and the limits of the reduced-form approach has led to the development of a new strand of literature based on quantitative models. Redding and Rossi-Hansberg (2017) have done a wonderful job in describing what has been accomplished since the beginning of this decade by using such models in spatial economics. So, we have little to add. According to us, one of the greatest merits of these models is to take on board the main ingredients of trade and regional and urban economics, while accounting for additional effects such as the heterogeneity of an arbitrary number of regions or cities. Locations are differentiated exogenously by given inhomogeneities in productivity and/or amenities and endogenously by the size of the local population. In other words, quantitative models may be viewed as sound and large comparative statics exercises that check the robustness of results obtained in "toy-models" studied in Section 3, which also share several features of the LUTI models discussed in section 3.3.

Quantitative models are also useful to study how the internal structure of a city is affected by various shocks. Along these lines, what is probably the best illustration of what has been achieved is the setting developed by Ahlfeldt *et al.* (2015), which is both tractable and amenable to empirical analysis. Two results are worth stressing. First, Ahlfeldt *et al.* find that substantial differences in productivity and amenities across locations are determined by the concentration of activities. Second, the elasticity of productivity with respect to density within cities is slightly higher than those reported in across-cities estimations.

Finally, quantitative models allow addressing new questions. For example, as seen above, little is known 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. Less expected, Behrens *et al.* also find that neither type of frictions significantly affects the U.S. city-size distribution. This suggests that changing spatial frictions affects the relative size of cities but not much their locations and ranks in the urban hierarchy, as in Desmet and Rossi-Hansberg (2013). This suggests that the overall level of spatial frictions may be less important that the relative levels. General improvements in transport technologies benefit all places more or less equally. However, new local infrastructure benefits some places much more than others. This may have large effects.

Note also that many quantitative models rely on the assumption of homogeneous and perfectly mobile workers. That so many people live in places where jobs are not available suggests that individuals have different attitudes toward mobility. This has led Diamonds (2016) and Redding (2016) to describe individual mobility by means of a logit-like discrete choice model. As a result, particular workers are sorted out according to their productivity whereas others will be gathered along their preferences for amenities.

The pros and cons of the quantitative models are well known (Holmes, 2010). In our opinion, one of their main pitfalls is the repeated use of the same functional forms. Admittedly, this facilitates comparisons. However, we do not learn anything about the robustness of the results. We have seen above that the iceberg cost function provides a relatively bad approximation of what actual transport rates are. This is an issue in models that aim to study the spatial interdependence among regions or cities through trade frictions. Furthermore, CES preferences are almost ubiquitous in spatial quantitative economics. As models are exactly calibrated, a poor choice of functional specifications passes easily unnoticed. The CES combined with the iceberg cost leads to a wide range of convenient properties which are, unfortunately, difficult to generalize (Parenti *et al.*, 2017). For example, as seen in 2.3 and 3.1, low transport costs intensify competition with the rest of the economy, while shocks to transport costs induce a tougher firm selection. The CES model of monopolistic competition cannot account for these effects. In addition, it is commonplace to use an upper-tier Cobb-Douglas utility nesting CES lower-tier utilities. In this case, the different sectors of the economy interact only through the local labor markets and the spatial distribution of regional incomes. This puts some severe limits on the general equilibrium effects that are typically associated with the existence of several sectors.

More problematically, the CES yields the socially optimal selection and firm-level outputs, but these two properties cease to hold for non-CES additive preferences (Dhingra and Morrow, 2017), which casts some doubt on the welfare assessments of policies and counterfactuals undertaken with such a social welfare function. Behrens *et al.* (2016) is a telling example of the possible biases associated with the CES. They consider a multi-sector model of monopolistic competition with CES and 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 $6-8\%$ under CARA preferences, while the welfare loss is much smaller under CES preferences (less than 1%).

So, if a structural analysis is highly desirable, it is not unreasonable to question its validly when it is applied to a very specific model. Not that long ago, there was a need to bring firms into standard trade theory. With the development of the new trade theories, a lot has been accomplished. Today, there is a need to bring heterogeneous consumers/workers into the picture. Given the on-going advances in numerical analysis, it should not be too hard to test the robustness of what we have learned by using models that have a richer demand side and a better description of the transport sector.

Concluding remarks

Regional and urban models are not full-fledged general equilibrium models à la Arrow-Debreu, nor are they Marshallian partial equilibrium models of a specific market. They combine a few particular markets, such as transport, land or labor, with the aim of studying their interaction while the rest of the economy provides the numéraire. The complexity of the issues at stake, the difference in geographical scales, the market structure problem, and the multiple facets of the transport sector explain why we need different models, as well as different econometric approaches. We have seen that several important questions remain open. Nevertheless, this survey aims to show that spatial economics is no longer a collection of unrelated concepts and methods, but a coalescing and growing field that has a large number of potential links with other economic fields, which are still unexplored.

Can spatial economics inform policy-makers? The answer is yes. Owing to the strength of market forces shaping the spatial economy, regional development seems to be inevitably unequal. To some extent, the unevenness of regional development may be viewed as the geographical counterpart of economic growth (Lucas, 1988). The cumulative nature of the agglomeration process makes the resulting pattern of activities particularly robust to various types of shocks. A spray-gun distribution of increasing-returns activities results in high investment expenditure and/or underutilization of infrastructure and facilities. Spatial dispersion of public investments is often inefficient because it prevents activities from reaching the critical mass needed to be efficient enough to compete on the national or international marketplace. Many policies fail to recognize that regional income differences are often the result of scale economies.

The historical and social background of a region, its economic strengths and weaknesses, its education system, its portfolio of amenities, and the quality of its governance are the fundamental ingredients to be accounted for when designing local development policies. Very much like firms differentiate their products to relax competition, regions must avoid head-to-head fiscal competition with well-established areas. Instead, regional development strategies should identify areas of specialization that exploit local sources of uniqueness. The aim of such strategies is to strengthen regions' comparative advantages and to give priority to finding sustainable solutions to regions' weakest links. Since firms have specific needs that vary with the kind of activity in which they are involved, a promising policy for locales is to design differentiated infrastructures. The scope for such a strategy is increasing as the revolution in information and communication technology has shifted firms' needs toward more specialized inputs.

Inequality cuts through the urban system. If anything else, the development of human capital should be the main target of urban policies. Rather than spending billions of dollars on large infrastructures and fancy buildings, local governments should promote the supply of affordable housing through the adoption of market-savvy land and construction policies and facilitate movement in cities by means of congestion pricing. Housing and transport markets are also intimately intertwined with local labor markets. Therefore, national employment policies that ignore the urban environment in which jobs are created are likely to be unable to deliver their full potential.

In the same vein, understanding how the process of land capitalization works might help finance local public goods, thus alleviating macroeconomic fiscal constraints. The land rent value at any specific location capitalizes (at least to a certain extent) the various costs and benefits generated in the vicinity of this location and its value is created by the actions taken by firms, households, and local governments, but not much by the landlords. Therefore, allowing landlords to capture the land rent is an implicit transfer from the collectivity to the landlords. Regional and urban policies informed by spatial economic insights can make national economies work better.

References

[1] Ahlfeldt, Gabriel, Stephen J. Redding, Daniel Sturm, and Klaus Wolf. 2015. The economics of density: Evidence from the Berlin Wall. *Econometrica* 83(6): 2127-89.

[2] 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 & Control* 36(11): 1729-59.

[3] Allen, Treb and Costas Arkolakis. 2014. Trade and the topography of the spatial economy. *Quarterly Journal of Economics* 129(3): 1085-140.

[4] Alonso, William. 1964. *Location and Land Use*. Cambridge (MA): Harvard University Press.

[5] 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.

[6] 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.

[7] Anderson, James E. and Eric van Wincoop. 2004. Trade costs. *Journal of Economic Literature* 42(3): 691-751.

[8] 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.

[9] 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.

[10] Antràs, Pol and Stephen R. Yeaple. 2014. Multinational firms and the structure of international trade. In: *Handbook of International Economics. Volume 4*, edited by Gita Gopinath, Elhanan Helpman and Kenneth Rogoff, 55-130. Amsterdam: Elsevier.

[11] Arnott, Richard. 2007. Congestion tolling with agglomeration externalities. *Journal of Urban Economics* 62(2): 187-203.

[12] 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.

[13] Arnott, Richard and Kenneth A. Small. 1994. The economics of traffic congestion. *American Scientist* 82: 446-55.

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

[15] Arrow, Kenneth J. and Gérard Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22(3): 265-90.

[16] Bacolod, Marigee, Bernardo S. Bull and William C. Strange. 2009. Skills in the city. *Journal of Urban Economics* 65(2): 136-53.

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

[18] Baldwin, Richard, Rikard Forslid, Philippe Martin, Gianmarco Ottaviano and Frédéric Robert-Nicoud. 2003. *Economic Geography and Public Policy*. Princeton (NJ): Princeton University Press.

[19] Baum-Snow, Nathaniel. 2007. Did highways cause suburbanization? *Quarterly Journal of Economics* 122(2): 775–805.

[20] Beckmann, Martin J. 1972. Spatial Cournot oligopoly. *Papers and Proceedings of the Regional Science Association* 28: 37-47.

[20] Beckmann, Martin J. 1976. Spatial equilibrium in the dispersed city. In *Mathematical Land Use Theory*, edited Yorgo Papageorgiou,117-25. Lexington (MA): Lexington Books.

[21] Behrens, Kristian, Gilles Duranton and Frédéric Robert-Nicoud. 2014. Productive cities: Sorting, selection and agglomeration. *Journal of Political Economy* 122(3): 507-53.

[22] 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.

[23] Behrens, Kristian, Andrea R. Lamorgese, Gianmarco I.P. Ottaviano, and Takatoshi Tabuchi. 2007. Changes in transport and non-transport costs: Local vs global impacts in a spatial network. *Regional Science and Urban Economics* 37(6): 625-48.

[24] 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.

[25] Behrens, Kristian, Giordano Mion, Yasusada Murata, and Jens Suedekum. 2016. Distorted monopolistic competition. CEPR Discussion Paper No.11642.

[26] Behrens, Kristian, Giordano Mion, Yasusada Murata, and Jens Suedekum. 2017. Spatial frictions. *Journal of Urban Economics* 97: 40-70.0

[27] Belenzon, Sharon and Mark Schankerman. 2013. Spreading the word: Geography, policy, and knowledge spillovers. *Review of Economics and Statistics*, 95(3): 884-903.

[28] 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.

[29] Berliant, Marcus, Shinkun Peng and Ping Wang. 2002. Production externalities and urban configuration. *Journal of Economic Theory* 104(2): 275-303.

[30] Berry, Steven and Joel Waldfogel. 2010. Product quality and market size. *Journal of Industrial Economics* 58(1): 1-31.

[31] Bleakley, Hoyt and Jeffrey Lin. 2012. Thick-market effects and churning in the labor market: Evidence from U.S. cities. *Journal of Urban Economics* 72(2-3): 87-103.

[32] 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.

[33] Bosquet, Clément and Henry Overman. 2016. Why does birthplace matter so much? Sorting, learning and geography. CEPR Discussion Paper No. 11085.

[34] Breinlich, Holger, Gianmarco Ottaviano and Jonathan Temple. 2014. Regional growth and regional decline. In: *Handbook of Economic Growth, Volume 2*, edited by Philippe Aghion and Steven Durlauf, 683-779. Amsterdam: Elsevier.

[35] Brinkman, Jeffrey C. 2016. Congestion, agglomeration, and the structure of cities. *Journal of Urban Economics* 94: 13-31.

[36] Brueckner, Jan K. 2000. Urban sprawl: Diagnosis and remedies. *International Review of Regional Science* 23(2): 160-71.

[37] Brueckner, Jan K., Jacques-François Thisse and Yves Zenou. 1999. Why is Central Paris rich and Downtown Detroit poor? An amenity-based theory. *European Economic Review* 43(1): 91-107.

[38] Buzard, Kristy, Gerald A. Carlino, Robert M. Hunt, Jake K. Carr and Tony E. Smith. 2015. Localized knowledge spillovers: Evidence from the agglomeration of American R&D labs and patent data. Federal Reserve Bank of Philadelphia.

[39] Cavailhès, Jean, Carl Gaigné, Takatoshi Tabuchi and Jacques-François Thisse. 2007. Trade and the structure of cities. *Journal of Urban Economics* 62: 383-404

[40] Chamberlin, Edward. 1933. *The Theory of Monopolistic Competition*. Cambridge (MA): Harvard University Press.

[41] 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: 457-90.

[42] 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.

[43] Charnoz, Pauline, Claire Lelarge and Corentin Trevien. 2016. Communication costs and the internal organization of multi-plant businesses: Evidence from the impact of the French high-speed rail. INSEE Working Paper G 2016 / 02.

[44] Cheshire, Paul C., Max Nathan and Henry G. Overman. 2014. *Urban Economics and Urban Policy*. Cheltenham, U.K.: Edward Elgar.

[45] Ciccone, Antonio and Robert E. Hall. 1996. Productivity and the density of economic activity. *American Economic Review* 86(1): 54-70.

[46] Combes, Pierre-Philippe, Sylvie Démurger and Li Shi. 2015. Migration externalities in Chinese cities. *European Economic Review* 76: 152-67.

[47] Combes, Pierre-Philippe, Gilles Duranton and Laurent Gobillon. 2008. Spatial wage disparities: Sorting matters! *Journal of Urban Economics* 63(2): 723-42.

[48] 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.

[49] Combes, Pierre-Philippe and Laurent Gobillon. 2015. The empirics of agglomeration economies. In: *Handbook of Regional and Urban Economics. Volume 5*, edited by Gilles Duranton, J. Vernon Henderson and William C. Strange, 247-348. Amsterdam: Elsevier.

[50] 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.

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

[52] Davis, Donald R. and Jonathan Dingel. 2017. A spatial knowledge economy. Columbia University.

[53] 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.

[54] 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.

[55] De Borger, Bruno and Stef Proost. 2012. A political economy model of road pricing. *Journal of Urban Economics* 71: 79-92.

[56] De la Roca, Jorge and Diego Puga. 2017. Learning by working in big cities. *Review of Economics Studies* 84(1): 106-42.

[57] de Palma, André, Victor Ginsburgh, Yorgo Y. Papageorgiou and Jacques-François Thisse. 1985. The principle of minimum differentiation holds under sufficient heterogeneity. *Econometrica* 53(4): 767-81.

[58] 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.

[59] 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.

[60] Dixit, Avinash K. and Joseph E. Stiglitz. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67(3): 297-308.

[61] Desmet, Klaus and Esteban Rossi-Hansberg. 2014. Spatial development. *American Economic Review* 104(4): 1211-43.

[62] Dhingra, Swati and John Morrow. 2017. Monopolistic competition and optimum product diversity under firm heterogeneity. *Journal of Political Economy*, forthcoming.

[63] Donaldson, Dave. 2017. Railroads of the Raj: Estimating the impact of transportation infrastructure. *American Economic Review*, forthcoming.

[64] Donaldson, Dave and Richard Hornbeck. 2016. Railroads and American economic growth: A "market access" approach. *Quarterly Journal of Economics* 131(2): 799-858.

[65] Duranton, Gilles and Diego Puga. 2004. Micro-foundations of urban increasing returns: Theory. In: *Handbook of Regional and Urban Economics. Volume 4*, edited by J. Vernon Henderson and Jacques-François Thisse, 2063-117. Amsterdam: Elsevier.

[66] Duranton, Gilles and Matthew A. Turner. 2011. The fundamental law of road congestion: Evidence from US cities. *American Economic Review* 101(6): 2616-52.

[67] Duranton, Gilles and Matthew A. Turner. 2012. Urban growth and transportation. *Review of Economic Studies* 79(4): 1407-40.

[68] Eaton, Jonathan and Zvi Eckstein. 1997. Cities and growth: Theory and evidence from France and Japan. *Regional Science and Urban Economics* 27(4-5): 443–74.

[69] Eeckhout, Jan, Roberto Pinheiro and Kurt Schmidheiny. 2014. Spatial sorting. *Journal of Political Economy* 122(3): 554-620.

[70] Fabber, Benjamin. 2014. Trade integration, market size, and industrialization: Evidence from China's national trunk highway system. *Review of Economic Studies* 81(3): 1046-70.

[71] Faggio, Giulia, Olmo Silva and William C. Strange. 2017. Heterogeneous agglomeration. *Review of Economics and Statistics*, forthcoming.

[72] Feldman, Maryann P. and Dieter F. Kogler. 2010. Stylized facts in the geography of innovation. In: *Handbook of the Economics of Innovation. Volume 1* edited by Bronwyn H. Hall and Nathan Rosenberg, 381-410. Amsterdam: Elsevier.

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

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

[75] 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.

[76] Fujita, Masahisa and Jacques-François Thisse. 2013. *Economics of Agglomeration. Cities, Industrial Location and Globalization*. Second edition. Cambridge: Cambridge University Press.

[77] García-López, Miquel-Ángel, Adelheid Hollb and Elisabet Viladecans-Marsal. 2015. Suburbanization and highways: When the Romans, the Bourbons and first automobiles still shape Spanish cities. *Journal of Urban Economics* 85: 52-67.

[78] Gaspar, Jess and Edward L. Glaeser. 1998. Information technology and the future of cities. *Journal of Urban Economics* 43: 136-56.

[79] 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(3): 745-63.

[80] Giroud, Xavier. 2013. Proximity and investment: Evidence from plant-level data. *Quarterly Journal of Economics* 128(2): 861-915.

[81] Glaeser, Edward L. 1999. Learning in cities. *Journal of Urban Economics* 46(2): 254-77.

[82] Glaeser, Edward L. 2011. *Triumph of the City*. London: Macmillan.

[83] Glaeser, Edward L., Joseph Gyourko and Raven E. Saks. 2006. Urban growth and housing supply. *Journal of Economic Geography* 6(1): 71-89.

[84] Glaeser, Edward L., Hedi D. Kallal, Jose Scheinkman and Andrei Shleifer. 1992. Growth in cities. *Journal of Political Economy* 100(6): 1126-52.

[85] Glaeser, Edward L. and Janet E. Kohlhase. 2004. Cities, regions and the decline of transport costs. *Papers in Regional Science* 83(1): 197-228.

[86] Glaeser, Edward L., Jed Kolko and Albert Saiz. 2001. Consumer city. *Journal of Economic Geography* 1(1): 27-50.

[87] Glaeser, Edward L. and David C. Maré. 2001. Cities and skills. *Journal of Labor Economics* 19(2): 316-42.

[88] Head, Keith and Thierry Mayer. 2004. The empirics of agglomeration and trade. In: *Handbook of Regional and Urban Economics. Volume 4,* edited by J. Vernon Henderson and Jacques-François Thisse, 2609-69. Amsterdam: Elsevier.

[89] Head, Keith and Thierry Mayer. 2014. Gravity equations: Workhorse, toolkit, and cookbook. In: *Handbook of International Economics. Volume 4*, edited by Gita Gopinath, Elhanan Helpman and Kenneth Rogoff, 131-96. Amsterdam: Elsevier.

[90] 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. Cambridge: Cambridge University Press, 33-54.

[91] Henderson, J. Vernon. 1974. The sizes and types of cities. *American Economic Review* 64(4): 640-56.

[92] Henderson, J. Vernon. 1988. *Urban Development. Theory, Fact and Illusion.* Oxford: Oxford University Press.

[93] Henderson, J. Vernon, Ari Kuncoro and Matt Turner. 1995. Industrial development in cities. *Journal of Political Economy* 103(5): 1067-90.

[94] Henderson, J. Vernon and Arindam Mitra. 1996. The new urban landscape: Developers and edge cities. *Regional Science and Urban Economics* 26(6): 613-43.

[95] Henderson, J. Vernon and Yukako Ono. 2008. Where do manufacturing firms locate their headquarters? *Journal of Urban Economics* 63(2): 431-50.

[96] Hohenberg, Paul and Lynn H. Lees. 1993. *The Making of Urban Europe (1000-1950).* Cambridge (MA): Harvard University Press.

[97] Holmes, Thomas, J. 2010. Structural, experimentalist, and descriptive approaches to empirical work in regional economics. *Journal of Regional Science* 50(1): 6-22.

[98] Hoover, Edgar M. 1936. *Location Theory and the Shoe and Leather Industries.* Cambridge (MA): Harvard University Press.

[99] Hoover, Edgar M. 1937. Spatial price discrimination. *Review of Economic Studies* 4(3): 182-91.

[100] Hotelling, Harold. 1929. Stability in competition. *Economic Journal* 39(153): 41-57.

[101] Hsieh, Chang-Tai and Enrico Moretti. 2015. Why do cities matter? Local growth and aggregate growth. NBER Working Paper No. 21154.

[102] Hurter, Arthur P. and Joseph S. Martinich. 1989. *Facility Location and the Theory of Production.* Dordrecht: Kluwer Academic Publishers.

[103] Hymel, Kent . 2009. Does traffic congestion reduce employment growth? *Journal of Urban Economics* 65: 127-35.

[104] Kaldor, Nicholas. 1935. Market imperfection and excess capacity. *Economica* 2(5): 35-50.

[105] Kalnins, Arturs and Francine Lafontaine. 2013. Too far away? The effect of distance to headquarters on business establishment performance. *American Economic Journal: Microeconomics* 5(3): 157-79.

[106] Knight, Brian. 2002. 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.

[107] Koskinen, Kaj U. and Hannu Vanharanta. 2002. The role of tacit knowledge in innovation processes of small technology companies. *International Journal of Production Economics* 80: 57-64.

[108] Krugman, Paul. 1980. Scale economies, product differentiation, and the pattern of trade. *American Economic Review* 70(5): 950-59.

[109] Krugman, Paul. 1991. Increasing returns and economic geography. *Journal of Political Economy* 99(3): 483-99.

[110] Krugman, Paul and Anthony J. Venables. 1995. Globalization and the inequality of nations. *Quarterly Journal of Economics* 110(4): 857-80.

[111] 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: 83-126.

[112] Leamer, Edward E. and Michael Storper. 2001. The economic geography of the Internet age. *Journal of International Business Studies* 32(4): 641-65.

[113] Lin,Yatang. 2017. Travel costs and urban specialization patterns: Evidence from China's high speed railway system. *Journal of Urban Economics* 98: 98-123.

[114] Lösch, August. 1940. *Die Räumliche Ordnung der Wirtschaft*. Jena: Gustav Fischer. English translation: *The Economics of Location*. New Haven (CN): Yale University Press (1954).

[115] Lucas Jr., Robert E. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22(1): 3-42.

[116] Lucas Jr., Robert E. and Esteban Rossi-Hansberg. 2002. On the internal structure of cities. *Econometrica* 70(4): 1445-76.

[117] Manning, Alan. 2010. The plant size-place effect: Agglomeration and monoposony power in labour markets. *Journal of Economic Geography* 10(5): 717-44.

[118] Martin, Philippe and Carol A. Rogers. 1995. Industrial location and public infrastructure. *Journal of International Economics* 39(3-4): 335-51.

[119] Marshall, Alfred. 1890. *Principles of Economics*. London: Macmillan. 8-th edition published in 1920.

[120] McCann, Philip. 2005. Transport costs and new economic geography. *Journal of Economic Geography* 5(5): 305-18.

[121] Mills, Edwin S. 1967. An aggregative model of resource allocation in a metropolitan area. *American Economic Review* 57(2): 197-210.

[122] Mion, Giordano and Paolo Naticchioni. 2009. The spatial sorting and matching of skills and firms. *Canadian Journal of Economics* 42(1): 28-55.

[123] Moretti, Enrico. 2012. *The New Geography of Jobs*. Boston: Houghton Mifflin Harcourt.

[124] Murata, Yasusada and Jacques-François Thisse. 2005. A simple model of economic geography à la Helpman-Tabuchi. *Journal of Urban Economics* 58(1): 137-55.

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

[126] Nitsch, Volker. 2005. Zipf zipped. *Journal of Urban Economics* 57: 86-100.

[127] Pigou, Arthur C. 1920. *The Economics of Welfare*. 4th ed. London: Macmillan.

[128] Ogawa, Hideaki and Masahisa Fujita. 1980. Equilibrium land use patterns in a nonmonocentric city. *Journal of Regional Science* 20(4): 455-75.

[129] O'Hara, Donald J. 1977. Location of firms within a square central business district. *Journal of Political Economy* 85(6): 1189-207.

[130] Ottaviano, Gianmarco I.P. and Jacques-François Thisse. 2004. Agglomeration and economic geography. In: *Handbook of Regional and Urban Economics. Volume 4*, edited by J. Vernon Henderson and Jacques-François Thisse, 2563-608. Amsterdam: Elsevier.

[131] Nocke, Volker. 2006. A gap for me: Entrepreneurs and entry. *Journal of the European Economic Association* 4(5): 929-55.

[132] Ohlin, Bertil. 1933. *Interregional and International Trade*. Cambridge, MA: Harvard University Press). Revised version published in 1968.

[133] Parenti, Mathieu, Philip Ushchev and Jacques-François Thisse. 2017. Toward a theory of monopolistic competition. *Journal of Economic Theory* 167: 86-115.

[134] Parry, Ian W.H and Kenneth A. Small. 2009. Should urban subsidies be reduced? *American Economic Review* 99(3): 700-24.

[135] Petersen, Mitchell A. and Raghuram G. Rajan. 2002. Does distance still matter? The information revolution in small business lending. *Journal of Finance* 57(6): 2533-70.

[136] Phlips, Louis. 1983. *The Economics of Price Discrimination.* Cambridge: Cambridge University Press.

[137] 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.

[138] Puga, Diego. 1999. The rise and fall of regional inequalities. *European Economic Review* 43(2): 303–34.

[139] Redding, Stephen. 2011. Economic geography: A review of the theoretical and empirical literature. In: *The Palgrave Handbook of International Trade*, edited by D. Bernhofen, R. Falvey, D. Greenaway and U. Kreickemeie. London: Palgrave Macmillan.

[140] Redding, Stephen J. 2016. Goods trade, factor mobility and welfare. *Journal of International Economics* 101: 148-67.

[141] Redding, Stephen J. and Esteban Rossi-Hansberg. 2017. Quantitative spatial economics. *Annual Review of Economics*, forthcoming.

[142] Redding, Stephen and Daniel Sturm. 2008. The cost of remoteness: Evidence from German division and reunification. *American Economic Review* 98(5): 1766-97.

[143] Redding, Stephen J. and Matthew A. Turner. 2015. Transportation costs and the spatial organization of economic activity. In: *Handbook of Regional and Urban Economics. Volume 5*, edited by Gilles Duranton, J. Vernon Henderson and William C. Strange 247-348. Amsterdam: Elsevier.

[144] Redding, Stephen and Anthony J. Venables. 2004. Economic geography and international inequality. *Journal of International Economics* 62: 53-82.

[145] Robert-Nicoud, Frédéric. 2005. The structure of simple 'New Economic Geography' models (or, On identical twins). *Journal of Economic Geography* 5(2): 201-34.

[146] Rosenthal, Stuart S. and William C. Strange. 2004. Evidence on the nature and sources of agglomeration economies. In: *Handbook of Regional and Urban Economics. Volume*

4, edited by J. Vernon Henderson and Jacques-François Thisse, 2119-71. Amsterdam: Elsevier.

[147] Rossi-Hansberg, Esteban. 2005. A spatial theory of trade. *American Economic Review* 95(5): 1464-91.

[148] Salop, Steven C. 1979. Monopolistic competition with an outside good. *Bell Journal of Economics* 10(1): 141-56.

[149] Schiff, Nathan. 2015. Cities and product variety: Evidence from restaurants. *Journal of Economic Geography* 15(6): 1085-123.

[150] 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.

[151] 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.

[152] Starrett, David. 1978. Market allocations of location choice in a model with free mobility. *Journal of Economic Theory* 17(1): 21-37.

[153] Syverson, Chad. 2004. Market structure and productivity: A concrete example. *Journal of Political Economy* 112(6): 1181-222.

[154] Tabuchi, Takatoshi and Jacques-François Thisse. 2002. Taste heterogeneity, labor mobility and economic geography. *Journal of Development Economics* 69(1): 155-77.

[155] Tabuchi, Takatoshi, Jacques-François Thisse and Xiwei Zhu. 2018. Does technological progress magnify regional disparities? *International Economic Review*, forthcoming.

[156] Takahashi, Toshiaki, Hajime Takatsuka and Dao-Zhi Zeng. 2013. Spatial inequality, globalization, and footloose capital. *Economic Theory* 53(1): 213-38.

[157] Thomas, Isabelle. 2002. *Transportation Networks and the Optimal Location of Human Activities: A Numerical Geography Approach.* Cheltenham, U.K.: Edward Elgar.

[158] Tombe, T. and X. Zhu. 2017. Trade, migration and productivity: A quantitative analysis of China. *American Economic Review*, forthcoming.

[159] 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.

[160] Vickrey, William S. 1964. *Microstatics.* New York: Harcourt, Brace and World.

[161] Vickrey, William S. 1969. Congestion theory and public investment. *Papers and Proceedings of the Eighty-first Annual Meeting of the American Economic Association*: 251-60.

[162] von Thünen, Johann H. 1826. *Der Isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. Hamburg: Perthes. English translation: *The Isolated State*. Oxford: Pergamon Press (1966).

[163] Winston, Clifford. 2013. On the performance of the US transport system: Caution ahead. *Journal of Economic Literature* 51(3): 773-824.

Contact details:
Jacques-François Thisse
CORE-Université catholique de Louvain (Belgium), Professor Emeritus.
National Research University Higher School of Economics (Moscow, Russia),
Center for Market Studies and Spatial Economics, chief research fellow.
E-mail: jacques.thisse@uclouvain.be, jthisse@hse.ru