DISCUSSION PAPER SERIES

No. 10790

SKILLED CITIES, REGIONAL DISPARITIES, AND EFFICIENT TRANSPORT: THE STATE OF THE ART AND A RESEARCH AGENDA

Stef Proost and Jacques-François Thisse

INTERNATIONAL TRADE AND REGIONAL ECONOMICS

Centre for Economic Policy Research

ISSN 0265-8003

SKILLED CITIES, REGIONAL DISPARITIES, AND EFFICIENT TRANSPORT: THE STATE OF THE ART AND A RESEARCH AGENDA

Stef Proost and Jacques-François Thisse

Discussion Paper No. 10790
August 2015
Submitted 17 August 2015

Centre for Economic Policy Research
77 Bastwick Street, London EC1V 3PZ, UK
Tel: (44 20) 7183 8801
www.cepr.org

SKILLED CITIES, REGIONAL DISPARITIES, AND EFFICIENT TRANSPORT: THE STATE OF THE ART AND A RESEARCH AGENDA[†]

Abstract

The three themes of this survey—cities, regions, and transport—are closely intertwined and gathered in the category R of the JEL Classification System. We discuss cities and regions in separate sections because they are different spatial units facing specific problems. Transport issues affect both cities and regions and are discussed in each relevant section. The introductory remarks explain both the reason for this division, as well as what spatial economics is all about. Because general economists have barely met the words cities, regions, and transport during their studies, we explain what the field of spatial economics is and define basic concepts that might not currently be in their tool box. The second section is devoted to cities; the third focuses on regions. We conclude with general policy recommendations.

JEL Codes: R00
Keywords: cities, land, congestion, region, trade, transport

Stef Proost Stef.Proost@kuleuven.be
CES-KU Leuven

Jacques-François Thisse jacques.thisse@uclouvain.be
CORE-UC Louvain

[†] We thank F. di Comite, M. Lafourcade, F. Robert-Nicoud, J. Südekum, A. Turrini, K. Van Dender, E. Viladecans-Marsal, as well as R. Lindsey, J-Cl. Prager, E. Quinet, K. Small, R. Evers and C-M. Fung, for their comments on previous drafts. We are grateful to M. Ivaldi, W. Leininger and M. Paasi for their encouragements and the organization of the COEURE and DG-REGIO workshops, and acknowledge the funding by the FP7/320300 grant under the FP7 SSH research program.

Contents

On ne fait point de l'industrie et du commerce entre ciel et terre ;
il faut se poser quelque part sur le sol.
Léon Walras, *Eléments d'économie politique pure*

1. What is spatial economics about?

The Industrial Revolution has exacerbated regional disparities by an order of magnitude that was unknown before. The recent development of new information and communication technologies is triggering a new regional divide that governments and the public opinion should be aware of. What are the economic tools that can be used to understand those evolutions? As spatial economics is about bringing location, distance, and land into economics, its aim is to explain where economic activities locate. This makes it is one of the main economic fields that can be used to understand how the new map of economic activities is been redrawing. Yet, at first glance, the steady, and actually spectacular, drop in transportation costs since the mid-nineteenth century—compounded by the decline of protectionism in the post-World War II era and, more recently, by the near-disappearance of communication costs—is said to have freed firms and households from the need to be located near one another. Therefore, it is tempting to foresee the "death of distance" and the emergence of a "flat world" in which competition should be thought of as a race to the bottom, with the lowest-wage countries as the winners.

But—and it is a big but—while it is true that the importance of proximity to natural resources has declined considerably, this does not mean that distance and location have disappeared from economic life. Quite the contrary is true. Recent work in regional and urban 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. Empirical evidence shows that sizable and lasting differences in income per capita and unemployment rates exist at very different spatial scales (country, region, city, and neighborhood). In brief, *the fundamental question of spatial economics is to explain the existence of peaks and valleys in the spatial distribution of wealth and people*. This is what we aim to accomplish in this survey, with a special emphasis on the role of large cities and transport policies. Most graduate or undergraduate students in economics have barely met the words cities, regions, and transportation during their studies.[1] We therefore will define all the basic concepts that are not part of the tool box of most economists. In particular, we show how the tools of modern economic theory can illuminate the main issues of spatial economics, and how modern empirical methods have helped measure them. Conversely, introducing space into economic modeling allows one to revisit existing theories and to suggest new

[1] In his well-documented and reader-friendly history of economic analysis, Sandmo (2011) discusses the work of von Thünen and Hotelling. However, the former is presented mainly as a forerunner of marginalism and the latter as a pioneer of oligopoly theory.

solutions to old problems. In particular, we highlight some of the findings that reveal the increased importance of space in the modern economy.

Many ideas and concepts have, admittedly, been around for a long time. However, they were fairly disparate and in search of a synthesis. To a large extent, the history of spatial economics may be viewed as a process that has gradually unified various bodies of knowledge within a theoretical framework in which the focus has shifted from perfect competition to imperfect competition and various types of market failures. The state of the art today is sufficiently advanced to sketch such a unified framework that could be the backbone of spatial economics (Fujita and Thisse, 2013).

Another point is worth making at the onset. Space is the substratum of human affairs, but space is also a consumption and production good in the form of land. Regional economics may be thought as "space without land," whereas urban economics is "space with land." The worldwide supply of land vastly exceeds the demand for land. As a consequence, the price of land should be zero. Yet, we all know that housing costs vary enormously with the size of cities for reasons that do not depend on the quality of the housing structure. Therefore, the price of land reflects the scarcity of "something" that differs from land per se.

1.1 Location does matter

Ever since the pioneering work of von Thünen (1826), a fundamental question has haunted spatial economics: Why do economic activities cluster in a few places? And why do cities exist at all? At first sight, the second question might be seen as a question for architects, engineers, or urban planners, not for economists. Indeed, until recently, economists have not paid much attention to cities as an economic object. This is probably because there is no satisfactory answer to that question within the dominant paradigm of economic theory, which combines perfect competition and constant returns to scale. In the absence of scale economies, fragmenting production into smaller units at different locations does not reduce the total output available from the same given inputs, but transportation costs decline. In the limit, if the distribution of natural resources is uniform, the economy is such that each individual produces for his or her own consumption. This strange world without cities has been called "backyard capitalism." To put it differently, each location would become an autarky, except it is possible that trade between locations might occur if the geographic distribution of natural resources is uneven. Admittedly, different locations do not a priori provide the same exogenous amenities. However, using the unevenness of natural resources as the only explanation for the existence of large cities and for regional imbalance seems weak. Rather, as noted by Koopmans (1957) more than 50 years ago, *increasing returns are critical to understanding how the space-economy is shaped*.

A simple example will illustrate this fundamental idea. Suppose a planner has to decide where to locate one or two facilities to provide a certain good to a population of immobile users who are evenly distributed between two

regions. Individual demands are perfectly inelastic and normalized to one; the marginal production cost is constant and normalized to zero. Consumers in the domestic region may be supplied at zero cost, whereas supplying those living in the foreign region entails a transportation cost of T euros. If two facilities are built, the cost of building a facility is equal to F euros in each region. If only one facility is made available, the planner must incur cost F; if two facilities are built, the cost is 2F. A planner who aims to minimize total costs will choose to build a facility in each region if, and only if, F + T is more than 2F, that is, T > F. This will hold when F is small, T is high, or both. Otherwise, it will be less expensive to build a single facility that supplies all people in both regions. In other words, weak increasing returns—F takes on high values—promote the scattering of activities, whereas strong increasing returns foster their spatial concentration. As a consequence, the intensity of increasing returns has a major implication for the spatial organization of the economy:

> **The first law of spatial economics:** ***If 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, a relatively small number of places account for a large share of the national value added, whereas many large areas account for none or little economic activity. The difficulty encountered by economists when they take into account scale economies in general equilibrium theory probably explains why spatial economics has been at the periphery of economics for so long.

That said, it must be kept in mind that accounting for increasing returns often yields a message that differs from the standard neoclassical paradigm of perfect competition and constant returns to scale. Even though transport costs must be positive for space to matter, one should not infer from this observation that location matters less when transport costs decrease—quite the opposite. Spatial economics shows that lower transport costs make firms more sensitive to minor differences between locations. To put it another way, *a tiny difference may have a big impact on the spatial distribution of economic activity*.

1.2 Moving goods and people is still costly

Transportation economics is not a subfield of spatial economics. However, *transportation cuts across both urban and regional economics*, but in a different way. Urban economics primarily focuses on the movement of people through their commuting and shopping behavior, whereas the shipping of commodities to spatial separated markets is a central issue in regional economics. Even though spatial economists rightly emphasize the impact of transport costs on the distribution of economic activities, it is fair to say that their modeling of the transportation sector is often simplistic. The cost of shipping goods is viewed as a hike in their marginal production costs, while commuting costs appear as an additional expenditure in consumers' budget. In addition, most economists disregard the role played by the transport sector in the development of particular cities or regions.

Ever since the beginning of the Industrial Revolution, there has been spectacular progress in terms of the speed and cost for urban, interregional, *and* international transport. According to Bairoch (1997), "overall, it can be estimated that, from 1800 to 1910, the decrease in (weighted) real average transport costs was on the order of 10 to 1." For the United States, Glaeser and Kohlhase (2004) observe that the average cost of moving a ton a mile in 1890 was 18.5 cents, as opposed to 2.3 cents today (in 2001 dollars). Yet, as will be seen, estimating the gravity equation reveals that distance remains a strong impediment to trade and exchange. What is more, the current concentration of people and activities in large cities and urban regions fosters steadily increasing congestion, both in private and public transport. Therefore, transportation faces different challenges at the urban and interregional levels. In the urban context, we concentrate mainly on commuting by means of different transport modes. In the regional context, transportation consists of interregional and international freight trips of inputs and outputs, as well as passenger trips.

In this paper, transportation refers to the movement of people, goods, information, or anything else across space. In the wonderful dimensionless world of some analysts, transportation costs are zero, and thus any agent is equally connected to, or globally competes with, any other agent. If the monetary cost of shipping goods has dramatically decreased, other costs related to the transport of goods remain significant. For example, the opportunity cost of time rises in a growing economy, so that the time cost wasted in shipping certain types of goods steadily rises. Similarly, doing business at a distance due to differences in business practices, as well as in political and legal climates, generates additional costs, even within the European Union (EU). Transportation costs still matter because the distance between locations affects the economic life under different disguises.

People move because they commute to their workplace, go shopping, drop their children off at schools, visit friends, and attend cultural events. Commuting is expensive and is also perceived by consumers as one of their most unpleasant activities. 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. These are big numbers and they confirm that commuting costs and traffic congestion are issues that are far too neglected by economists. At the interregional level, migration costs are still very high within the EU. Doing a back of the envelope calculation, Cheshire and Magrini (2006) find that, despite smaller regional disparities and larger average distances in the U.S. than in the EU, the net migration rate of a population of a given size between comparable areas (the 50 American States plus Washington, D.C. versus the EU-12 large countries—France, Germany, Spain and the UK—divided into their level 1 regions—in Germany the Länder—while the smaller countries are treated as single units) is almost 15 times higher in the U.S. than in the EU. These authors conclude that "in Europe, urban population growth seems likely to be a rather imperfect signal of changes in welfare in cities."

Even within European countries, migration is sluggish and governed by a wide range of intangible and time-persistent factors. For example, controlling for the geographical distance and several other plausible effects, Falck

et al. (2012) show that actual migration flows between 439 German districts (the NUTS 3 regions) are positively affected by the similarity of the dialects prevalent in the source and the destination areas more than 120 years ago. In the absence of such dialects, which are seldom used today, internal migration in Germany would be almost 20 per cent higher than what it is. Further evidence of the low mobility of workers is provided by Bosquet and Overman (2015). Using the British Household Panel Survey that involves 32,380 individuals from 1991 to 2009, these authors observe that 43.7 percent of workers only worked in the area where they were born. Among the unskilled workers, this share grows to 51.7 but drops to 30.5 percent for workers having a college degree. Such low lifetime mobility provides empirical evidence that migration costs are an important determinant of the space-economy. Furthermore, 44.3 percent of the panel retirees live where they were born, which reveals a high individual degree of attachment to their birthplace.

To sum up, the transport of (some) goods remains costly; consumers spend a high proportion of their income on housing, while many services used by firms and households are non-tradable. Moreover, we will see that proximity remains critical for the diffusion of some information. European people are sticky, which means that the model widely used in the U.S. to study urban and regional growth, which relies on the perfect mobility of people and the search for amenities, has very limited application within the EU, not to say within European countries. These facts have a major implication for the organization of the (European) economic space:

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

Combining the first and second laws of spatial economics leads 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 the costs generated by the transfer of people, goods, and information*****.**

We may thus already conclude that high (low) transportation costs promote the dispersion (agglomeration) of economic activities, while strong (weak) increasing returns act as an agglomeration (dispersion) force. This trade-off is valid on all spatial scales (city, region, country, continent), which makes it a valuable analytical tool.[2] We will return to this in the next two sections.

[2] This trade-off has been rediscovered several times. It goes back at least to Lösch (1940).

1.3 Spatial scales

Spatial economics focuses on different *spatial scales*, which each correspond to a different level of spatial aggregation, ranging from the local to the global through the urban and the national. Economists too often use different, and perhaps equally unclear, words interchangeably—words such as locations, regions, or places—without being aware that they often correspond to different spatial units. Worse: sometimes, they use the same model to explain the location of economic activity at various spatial scales. In doing so, they may draw implications that are applicable at a certain spatial scale but not at another.

At the city level, locations are disaggregated. The city has a spatial extension because economic agents consume land, which implies that consumers travel within the city. Therefore, an urban space is both the substratum of economic activity and a private good (land) that is traded among economic agents. The main objective of *urban economics* is to explain how cities—which are to be understood as metropolitan areas that extend beyond the core city limits—are organized; that is, why are jobs concentrated in a few employment centers and how are consumers spatially distributed within the city according to their incomes and preferences? Central to the workings of a city is the functioning of its land market, which allocates both economic agents and activities across space, and of the quality of the transportation infrastructures used by commuters and shoppers. Equally important are various types of social networks that operate within very short distances. For example, informational spillovers affect positively the productivity of the local R&D sector, whereas neighborhood effects are often critical to sustain criminal activities in particular urban districts. To understand cities, we must view them not simply as places in space but as systems of market and non-market interactions that are anchored.

At the *interregional level*, locations are aggregated into subnational units that are distant from each other. Regardless of what is meant by a region, the concept is useful if, and only if, a region is part of a broader network through which various types of interactions occur. In other words, any meaningful discussion of regional issues requires at least two regions in which economic decisions are made. Hence, space is the substratum of activities, but land is not a fundamental ingredient of regional economics. Furthermore, as repeatedly stressed by Ohlin (1933), if we do not want the analysis to be confined to trade theory, we must also account explicitly for the mobility of agents (firms and/or consumers) and for the existence of transportation costs in trading commodities. However, how well a region does also depends on the functioning of its local markets and institutions. The surge of new economic geography (NEG) has allowed one to re-think regional economics in accordance with Ohlin's recommendation by combining the trade of goods and the mobility of production factors. In NEG, a region is assumed to be dimensionless and described by a node in a transportation network. The objective of *regional economics* is then to study the distribution of activities across a regional system. Figure 1 shows the geographical distribution of the GDP per capita per NUTS 3 region in the EU. We note striking differences across countries but

also within countries. Understanding these differences and what policies make sense is one of the principal motivations for this survey.

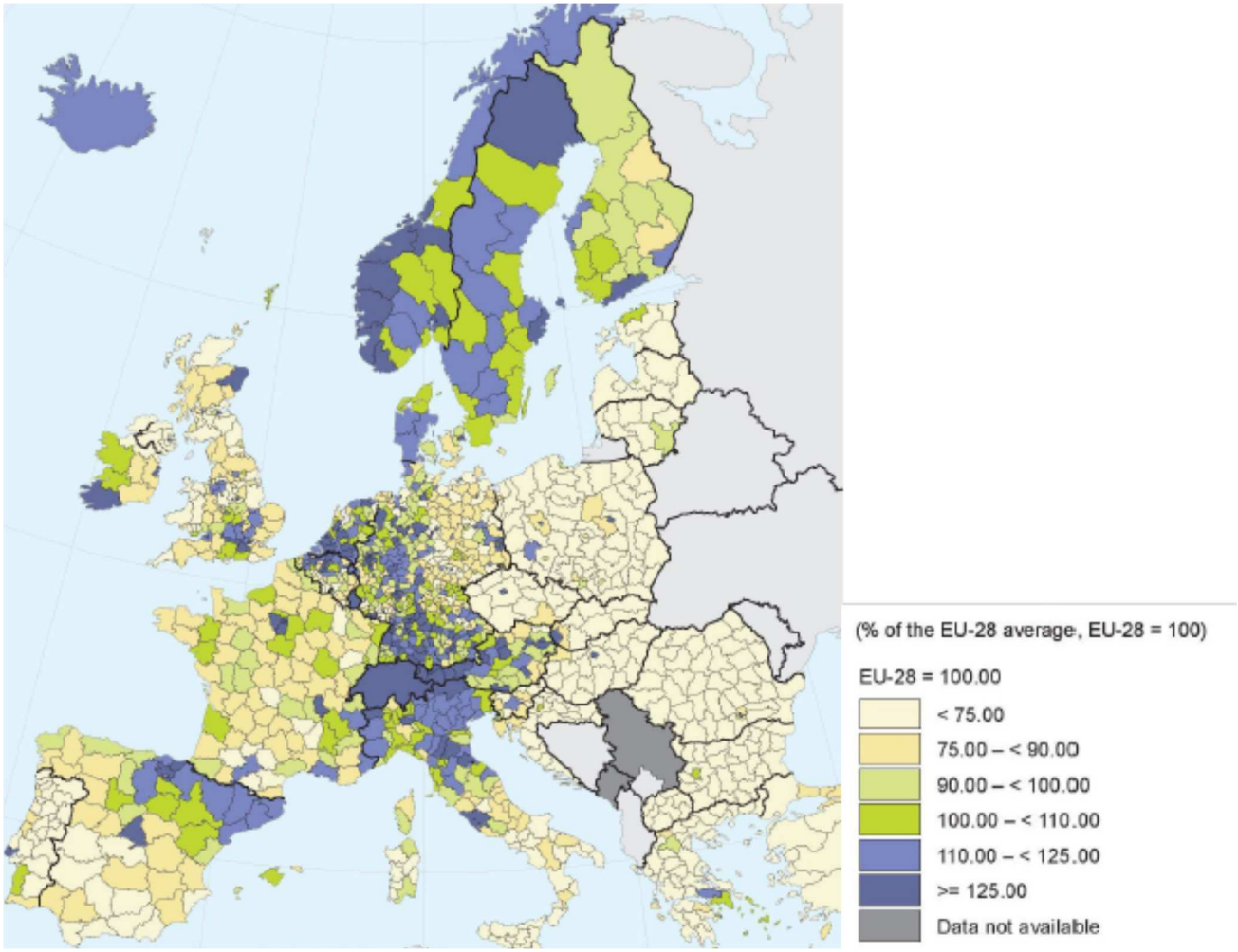

FIGURE 1. GDP per capita across NUTS 3 regions in 2011

Urban economics is a well-established field that draws on microeconomic concepts and tools. In contrast, the scientific status of regional economics is less clear as regional concepts, models, and techniques were too often merely extensions of those used at the national level, with an additional index identifying the different regions. The emphasis on standard trade theory in economic theory also hindered the further development of regional economics as trade is presented as a substitute for the mobility of factors. This point is especially well illustrated by the factor price equalization theorem.

Many prosperous regions are city-regions or regions accommodating a dense network of medium-sized cities such as the Randstad in the Netherlands. This idea is backed up by casual evidence: among the top 10 NUTS-2 regions

of the EU in terms of gross domestic product (GDP) per capita, 8 are formed by or organized around a major capital city. Figure 2 shows the range of the distribution of regional GDP per capita within each EU country; in most cases, the top position is occupied by the capital regions. In the U.S., the 20 largest metropolitan areas produce about half of the American GDP. This suggests that interregional systems should be studied in relation to urban systems.

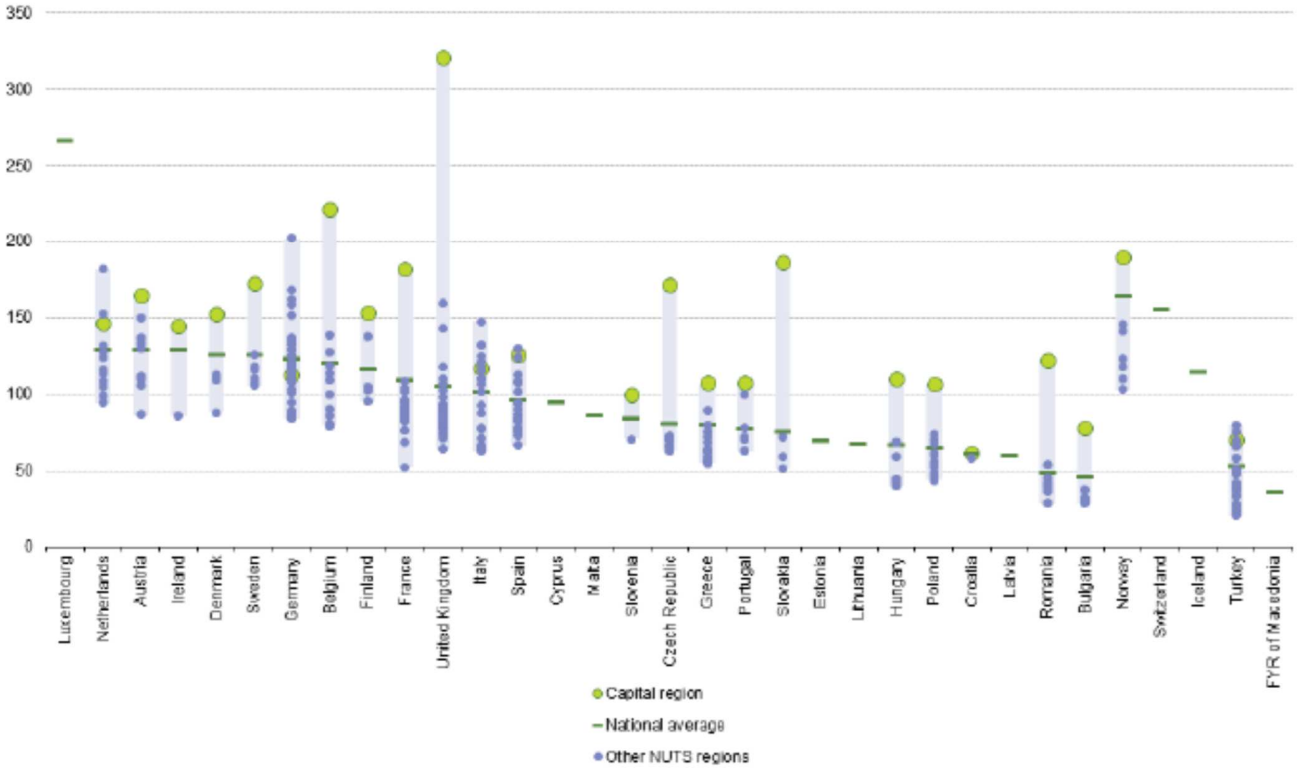

FIGURE 2. The distribution of the GDP per capita within EU countries

1.4 Are locations fixed or variable?

When the location of economic activity is given, the main issue is to determine the flows of goods across space, as well as the commuting and shopping trips undertaken by consumers. These flows are studied in two distinct fields, namely, trade theory and transportation economics. However, a full-fledged analysis of the space-economy has to be conducted within a framework in which firms and households choose their locations. Developing such models is especially worthwhile in a fast-evolving economic and technological environment in which firms and households face strong incentives to change location within and between cities, or across regions and countries.

At each spatial scale, economic areas are affected not only by the growing mobility of commodities but also by that of production factors (e.g., capital and labor). In particular, *lowering transport costs changes the nature and intensity of firms' and workers' incentives to move*. Therefore, to assess the full impact of trade and transport

policies, it is crucial to have a good understanding of how economic agents react to decreasing trade and transport costs**.** It is worth stressing that policy-makers often overlook the fact that their decisions affect the location choices made by firms and households.

Modern regional and urban economic theories highlight the fact that the rising mobility of goods and people need not reduce spatial inequality. In this respect, it is worth stressing that a distinctive feature of modern economic geographies is their *putty-clay* nature. Although many activities are a priori free to locate wherever they wish (putty), once they are anchored in particular cities or regions, the activities tend to grow there because agglomeration economies are localized (clay). Narrowing down interregional income gaps is the main objective of many regional policies and of the European Structural Funds, which accounts for the second largest share of the EU's budget. Although providing efficient transportation within large cities might be desirable because it facilitates movement in a dense network of relations, it is far less clear that the construction of big and expensive transportation infrastructures connecting regions delivers the expected effects. We will see that NEG suggests that a steady decrease in trade and transportation costs need not foster the geographical dispersion of activities. Why some regions fare better than others calls for explanations that go beyond the common wisdom of regional development agencies. We summarize the main characteristics of urban and regional economics in Table 1.

Table 1. Main characteristics of urban economics and regional economics

	Urban economics	**Regional economics**
Focus	Cities and metropolitan areas	EU, federal or large countries
Spatial scale	Location within cities	Location across regions
Capital mobility	Perfect	Imperfect/Perfect
Labor mobility	High via commuting	Low
Residential mobility	High	Low
Transport of goods	Trucks	Road, rail, water, air
Transport of passengers	Car, bus, metro, rail, bike, walk	Road, rail and HSR, air
Major issues	Agglomeration versus congestion	Global efficiency versus spatial equity

For a long time, spatial economics was confined to a small circle of specialists. One of the main features of modern spatial economics that makes it so exciting is its strong connection to other economic fields. First of all, spatial economics is now firmly rooted in microeconomics. Second, while regional economics is closely related to trade theory, it is also very much connected to modern industrial organization. In particular, the geography of a territory appears to be more and more dependent on the way firms organized their activities. Regarding urban economics,

since the early 1970s it has grown as a field of economics per se. However, it now has much stronger links to recent intellectual developments such as the new growth theories, where many scholars see cities as the engine of growth, as well as theories on social networks and other forms of local interactions, the urban neighborhood being the place where many nonmarket relationships are developed. Another distinctive feature of the recent contributions to regional and urban economics is the growing number of empirical studies, which aim at testing theories, using modern econometric methods. Transport studies have always been multidisciplinary. However, they remained for too long unrelated to regional and urban economics. This situation is changing fast and we may hope fruitful cross-fertilization.

2. Are cities a thing of the past?

The principal 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 on a daily basis for a variety of economic and social reasons. For example, individuals want to be close to each other because they like to meet other people, learn from others, and have a broader range of opportunities. Hence, *the main reason for the existence of cities is to connect people*. 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. Contrary to an opinion widely dispersed in the media, despite the Internet and other new communication devices, face-to-face contact remains important, at least for certain human and economic activities.[3] To understand why this is so, one must keep in mind that the information transferred by means of modern communication tools must be structured according to schemes and codes that are clearly defined and known to all. Only formal and precise information can be transmitted this way. In contrast, information that is difficult to codify can often be transmitted only through face-to-face contact. For example, the preliminary stages of developing a new technology or product require repeated contacts among those involved and such contacts are much easier and less costly when the people are in close proximity. As a result, cities are still the best locations for information-consuming activities, especially when firms operate in an environment of rapid technological change and strong competition.

Acquiring information is one of the main drivers of travel demand. Therefore, it follows that, once information is available at home through the Internet, the demand for travel should decrease. However, this argument overlooks the fact that, in a world where information is imperfect, people may refrain from visiting new destinations because there is too much uncertainty about what they can find there. In contrast, when information can be collected at low cost via the Internet or other devices, having more information may lead people to travel more often and to more distant locations because they have greater knowledge about the opportunities available in such locations.

[3] Frank (2014) uses experimental economics to shed new light on the importance of face-to-face contacts.

As a result, having more information may increase the demand for traveling, especially when traveling is cheap. Accordingly, Internet and transportation are both substitutes and complements, as were the telegraph and the telephone (Gaspar and Glaeser, 1998).

In the industrial era, cities allowed substantial decreases in transportation costs between large and connected production plants. Today, cities are the cradle of new ideas that benefit firms of very different sizes. This idea is not new, for cities are and have been for centuries the source of productivity gains as well as technological and cultural innovations (Hohenberg and Lees, 1985; Bairoch, 1985). To a large extent, it is fair to say that the agglomeration of economic activities is the geographical counterpart of social and economic development. However, these positive effects come with congestion, segregation, pollution, and crime. In addition, European cities are much older than American cities. The European cultural heritage is an advantage for economic and social development but is also a major constraint in terms of the organization and management of mobility within cities. This should not conflict with the awareness that wealth is increasingly created in cities, a fact that holds for the EU and, more generally, for developed and emerging countries alike. What is more, although there is not (yet) an urban strategy at the level of the EU, there is a growing recognition that many large European cities face similar social and cohesion problems.

A national economy is increasingly becoming a network of cities. Looking at cities through the lens of microeconomics sheds new light on issues that are otherwise often poorly understood. In what follows, we start by analyzing the fundamental forces that drive the formation and size of cities: (i) agglomeration economies generated by a dense web of activities and (ii) the trade-off between commuting and housing costs. Afterwards, we discuss more specific issues, with a special emphasis on urban transportation.

Many social scientists, including quite a few economists, would be skeptical about the idea of using microeconomics to study cities. We cannot find a better rebuttal to this objection than the following extract from Solow (1973, p.1):

> To study the locational equilibrium of a city seems almost silly. Buildings, streets, subways are among the most durable objects we make, and it is very expensive to move them or even to remove them. Existing patterns of location must therefore have been determined in large part by decisions that were made and events that happened under conditions that ruled long ago. It seems far-fetched to expect that what now exists will bear much relation to what would now be an equilibrium. Nevertheless, it turns out that the equilibrium states of simple models of urban location do actually reproduce some of the important characteristics of real cities. If this turns out to be more than mere coincidence, it is of some importance even for policy.

That said we may now proceed.

2.1 Agglomeration economies

Human beings have a pervasive drive to form and maintain lasting relations with others. Cities may thus be viewed, at least in the first order, as the outcome of the interplay between a field of social interactions and competition for land. Being isolated allows an individual to consume more land but renders interactions with others more costly. To study this trade-off, Beckmann (1976) assumes that the utility of an individual depends on the average distance to all individuals, as well as on the amount of land bought on the market. In equilibrium, the city exhibits a bell-shaped population density distribution supported by a similarly shaped land rent curve. In other words, *the natural gregariousness of human beings turns out to be sufficient to motivate them to gather within compact areas*. However, despite its relevance, such an explanation is not sufficient to explain the existence of urban agglomerations with millions of inhabitants.

It is well known that consumers in large metropolises pay high rents, have a longer commute, live in a polluted environment, and face high crime rates. So why would 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 larger cities' firms pay higher wages to their employees? If firms do not bear lower costs and/or earn higher revenues in large cities, they would rather locate in small towns or in the countryside where both land and labor are much cheaper. Why firms set up in large cities is now well documented: *the productivity of labor is higher in larger cities than in smaller ones*. Or, to put it bluntly, after controlling for unemployment and participation, *wages and employment* (both levels and rates) *move together*. This does not mean the demand for labor is upward-sloping. Instead, the reason for this urban wage premium is found in what economists call "agglomeration economies."

Whereas economists have long acknowledged the benefits associated with integrating international markets, it took them much longer to understand that similar benefits are associated with dense and thick markets such as those in large cities. Starting with the very 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 0.10 (Rosenthal and Strange, 2004). However, this result could be explained by the fact that some econometric problems have not been properly addressed.

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 amenities, is equivalent to assuming that skills or amenities 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. In doing so, we face the familiar quest of adding an endless string of control variables to the regressions.

Rather, using city/region and industry fixed effects allows us 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.[4]

Therefore, caution is needed when measuring the impact of employment density on labor productivity. Using advanced econometric methods and 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), urban economists have obtained more accurate estimations of agglomeration gains. There is now a broad consensus recognizing that, everything else being equal, the elasticity of labor productivity with respect to current employment density is about 0.03. This elasticity measures the static gains generated by a higher employment density. For example, doubling the employment density in Greater Paris would generate an increase in labor productivity that would be twice as large as what it would be in the least populated "départements" of France.

Dynamic gains, having a similar or slightly lower magnitude, are also at work the longer the individual stays in a denser area. In addition, the knowledge embodied in an individual who works in a dense area keeps producing its effects when this individual moves to a less efficient area. Therefore, the benefits generated by the agglomeration of activities are diffused through the spatial mobility of skilled workers.[5] Note, finally, that other variables, such as the share of skilled workers in the local labor force (Lucas, 1988) or the proximity to other productive and rich

[4] It is often easy to "rationalize" a correlation, or a chicken-and-egg problem, by a story showing that the causality runs in a specific direction. This issue, which we also encounter in many economic fields, pops up very often in urban and regional economics. The reader is refereed to Baum-Snow and Ferreira (2015) for a detailed survey.
[5] See Combes and Gobillon (2015) for a detailed survey.

areas, as stressed in NEG and discussed in Section 3, are also statistically significant (Combes et al., 2008). That said, what microeconomic effects stand behind agglomeration economies?

2.1.1 The nature and magnitude of agglomeration economies

We have seen that increasing returns are crucial to understanding the formation of the space-economy. The most natural way to think of increasing returns is when a plant having 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. Increasing returns may also materialize in 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*. In other words, even when individual firms operate under constant returns, there are increasing returns in the aggregate. In a nutshell, the whole is greater than the sum of its parts.

Duranton and Puga (2004) have proposed gathering the various effects associated with agglomeration economies into the following three categories: sharing, matching, and learning.

(i) ***Sharing*** primarily refers to local public goods provided to consumers and/or producers. When seeking a reason for the existence of cities, the one that comes most naturally to mind is the variety and quality of public services, as well as the availability of efficient and big infrastructures. This includes 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 also public services that increase consumers' well-being. A large number of people and firms 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. In other words, the efficiency of many public services rises when they are supplied to a dense population of users.[6] But sharing also refers to the supply of intermediate or business-to-business services, which are available in large markets. 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, that is, in big cities.

(ii) ***Matching*** means that the quality of matches between workers and firms on the labor market is higher in a thick market than in a thin one because of the larger number of opportunities agents face when operating in a denser labor market. How strong this effect is remains to be determined (Figueiredo et al., w2014). However, sticky workers living in small cities operate in markets with few potential employers, thereby allowing firms to

[6] For example, Gobillon and Milcent (2013) study the performance of French hospitals and find that a higher number of patients in a few large hospitals foster a lower mortality rate.

exploit their monopsony power and to pay lower wage.[7] In contrast, workers living in large cities do not have to change place to have a large array of potential employers. This makes the workers more prone to change jobs. Consequently, workers having the same individual characteristics will earn higher wages in larger cities than in smaller because firms have less monopsony power in thicker labor markets than in thinner (Manning, 2010).

(iii) ***Learning*** asserts that different agents own different bits of information so getting the agents together allows informational spillovers that raise the level of knowledge, thus improving firms' and workers' productivity. Spillovers stem from specific features of knowledge; in particular, knowledge is a non-rivalrous and partially excludable good. The role of information in modern cities has long been emphasized by economic historians. In the words of Hohenberg and Lees (1985), "urban development is a dynamic process whose driving force is the ability to put information to work. After 1850, the large cities became the nurseries as well as the chief beneficiaries of an explosion in knowledge-centered economic growth." Cities are the places where people talk. Undoubtedly, much of this talk does not generate productivity gains. However, the greater the number of people, the more likely the talk will lead to innovations, increasing productivity.

This is nothing new. In the 15th and 16th centuries, painting was one of the "high-tech" activities of the time. In *Lives of the Most Excellent Painters, Sculptors, and Architects*, Vasari writes the following: "It is a habit of Nature when she makes one man very great in any art, not to make him alone, but at the same time and in the same place to produce another to rival him, that they may aid each other by emulation." The congruence of talents is rarely natural. Rather, it is generally due to the attractive power of prosperous cities, such as Venice, Florence, Antwerp, and Amsterdam in the past, and New York, London, San Francisco, and Shanghai today. As Glaeser (2011) noted, "cities, and the face-to-face interactions that they engender, are tools for reducing the complex-communication curse."

Education generates an externality—the knowledge spillovers from skilled workers to other skilled workers—that did not attract much attention until recently (Lucas, 1988). Moretti (2004) has convincingly argued that the social productivity of education exceeds its private productivity. In other words, acquiring human capital enhances not only the productivity of the worker who acquires it but also the productivity of others because we learn from the others. What is important for the economic performance of cities is that skilled workers seem to benefit more from the presence of other skilled workers than the unskilled workers. Evidently, this effect is stronger in the case of regular, easy contacts between skilled workers. For example, Charlot and Duranton (2004) find that, in larger and more educated cities, workers exchange more than in cities populated by less-skilled workers. These authors show that such communications explain between 13 and 22 percent of the urban premium paid to the high-skilled

[7] This, and the low value of land, explains why many manufacturing firms have relocated their production plants from large to small cities.

workers. In the same spirit, Bacolod et al. (2009) observe that the urban wage premium associated with large cities stems from cognitive skills rather than motor skills. Therefore, 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. It is no surprise, therefore, that specific workers tend to sort out across space according to their skills.

In the U.S., Moretti (2012) observed that college graduates living in the richest cities, which are typically knowledge-based metropolitan areas, earn wages that are 50 percent higher than college graduates living in the bottom group of cities. In France, about half of spatial income disparities are explained by the different locations of skilled and unskilled workers (Combes et al., 2008), while between 85 and 88 percent of spatial wage disparities in the UK are explained by individual characteristics (Gibbons et al., 2014). The concentration of human capital and of high-value activities in large cities is a marked feature of developed and emerging economies. In other words, *spatial inequalities tend more and more to reflect differences in the distribution of skills and human capital across space.*[8] This has significant implications for the organization of the space-economy: cities specializing in high-tech activities attract highly skilled workers, who in turn help make these places more successful through other agglomeration economies and better amenities (Diamond, 2015). In other words, *workers tend to make diverging location choices by skills*. The downside of the spatial sorting of human capital across cities is the existence of stagnating or declining areas that specialized in industries with a limited human capital base, which are associated with low wages and a small number of local consumer businesses.

To a large extent, this evolution is enabled by the low transportation and communication costs prevailing today. Although these reduced costs allow standardized activities to be located in remote, low-wage countries, big cities remain very attractive to those activities where access to information and advanced technologies is of prime importance. Firms operating in industries that undergo rapid technological changes must be able to react quickly to market signals and to design specialized and sophisticated products that require a skilled labor force, especially when competition is intensified by low transportation costs. In a knowledge-based economy where information moves at an increasingly rapid pace, the value of knowledge and information keeps rising. Eventually, this increases the need for proximity for activities involving firms' strategic divisions, such as management, marketing, finance, and R&D, as well as specialized business-to-business (advertising, legal, and accounting services) and high-tech industries.[9]

[8] See, e.g. Glaeser and Maré (2001) and Moretti (2012) for the U.S., Combes et al. (2008) for France, Mion and Naticchioni (2009) for Italy, and Groot et al. (2014) for the Netherlands.

[9] Agglomeration effects may come in unexpected disguises: a firm, located in the vicinity of another firm belonging to the same sector and exporting to country A, has a higher probability to export to the same country (Koenig et al., 2010).

If the existence of informational and knowledge spillovers is indisputable, how to measure their magnitude is hard as they are not observable. Different strategies have been proposed to figure out what their importance is. One of the most original approaches is that of Arzaghi and Henderson (2008) who study the advertising agencies in Manhattan to infer networking effects among geographically close agencies. Advertising is an industry in which creativity matters a lot and where new ideas are quickly obsolete. Arzaghi and Henderson find that there is an extremely rapid spatial decay in the benefits of having close neighbors. They also show that firms providing high quality services locate close to other high-quality firms because they do not want to waste resources on discovering what neighbors have valuable information or establishing communication links with low-quality firms.

At the other extreme of the spectrum, a large-scale natural experiment has been conducted in Ctrip, China's largest travel agency, with 16,000 employees and a NASDAQ listing. The aim was to assess the impact of home working. Employees who volunteered to work from home were randomized by even/odd birthdate into a treatment group who worked from home four days a week for nine months and a control group who were in the office all five days of the work week. Bloom al. (2015) report that home working led to a 13 percent performance increase for those who worked home, and no negative spillovers onto workers who stayed in the office. It should be stressed, however, that unlike advertising agencies call centers are particularly suitable for telework as their activities requite neither teamwork nor face-to-face contacts. As expected, some activities do not benefit from agglomeration economies, which incite firms to conduct some of their activities in front offices located in the central city, whereas the rest of their activities are carried out in back offices set up in the suburbs or even in remote areas (Ota and Fujita, 1993).

All the agglomeration effects discussed above may be *intrasectoral* as in Marshall (1890) or *intersectoral* as in Jacobs (1969). Regardless of their origin, those effects have the nature of increasing returns *external* to firms. Recent empirical works show that their existence is unquestionable. However, as suggested by the above-mentioned examples, several issues remain unclear (Puga, 2010). First, different industries agglomerate for different reasons. Therefore, what is the relative importance of the various types of agglomeration economies in cities that specialize in different activities? Second, are agglomeration economies stronger in high-tech industries than in traditional sectors, which are typically less information-based? Third, the geographical distribution of human capital explains a large share of spatial inequalities. However, it is not clear how much of the effect of human capital is explained by the distribution of individual workers and by the presence of human capital externalities across highly skilled workers. Last, how does city size affect the nature and magnitude of agglomeration economies? For example, in a specialized city, a negative shock to the corresponding industry affects its workers negatively. In contrast, in a city endowed with a portfolio of industries, workers may expect to find a job in firms belonging to other industries. In other words, a diversified—and probably large—city acts as an insurance device. For example, large French cities have been less affected by the Great Recession than have other

territories (INSSE Première n°1503). In the same vein, unplanned interactions allow firms belonging to one sector to benefit from the presence of another located in the same city.

In a recent comprehensive study, Faggio et al. (2015) 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, but 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, if we want to design more effective policies for city development and redevelopment we need a deeper understanding of the drivers that stand behind the process of agglomeration. Furthermore, the interactions across agents are driven by the accessibility of an agent to the others. Although geographers and transportation economists consider the employment density is a rather crude proxy for accessibility, how to define and measure this one in econometric studies of agglomeration economies remains an open question.

2.1.2 Cities as consumption centers

The usual cliché is that big cities are bad for consumers. But the authors of anti-city pamphlets forget two things: (i) all over the world, free people vote with their feet by moving to cities; and (ii) cities are not just efficient production centers but are also great consumption, culture, and leisure places (Glaeser et al., 2001). Consumers living in large cities enjoy a wider range of goods, services, and contacts as the number of shops, cultural amenities, and opportunities for social relations all increase with city size. Even if dating tends to be more and more via the Internet, the two parties have to meet physically one day or another. While the steady decline in transportation costs and the progressive dismantling of tariff barriers have vastly improved the access to foreign goods, models in industrial organization show that the concomitant increase in competition incentivizes both incumbent and new firms to restore their profit margins by supplying higher-quality goods as well as more differentiated products. Because the taste and income ranges are greater in bigger cities that also allow more varieties to cover their fixed production costs, more goods and services are available in such markets (Picard and Okubo, 2012). In sum, consumers living in larger cities enjoy a broader range of goods and business-to-consumers services.

Even though, as shown below, housing is more expensive in large cities than in small, tradable goods need not be more expensive. Since a larger city provides a larger outlet for consumption goods, there is more entry, which intensifies competition; such cities attract the most efficient retailers that also benefit from agglomeration economies and better logistics. Again, as suggested by industrial organization theory, market prices tend to be lower in larger cities than in smaller and the number of varieties of a base product is higher. Calculating the first theoretically based urban price index for 49 U.S. cities, Handbury and Weinstein (2015) show that prices will fall by 1.1 percent when population doubles, while the number of available products will increase by 20 percent.

These consumption benefits become even more pronounced once it is recognized that the hierarchy of public services is often the mirror image of the urban hierarchy. In particular, the congregation of a large number of people facilitates the mutual provision of public services that could not be obtained in isolation. Health care and educational facilities are good cases in point.

Notwithstanding many qualifications, the empirical evidence in the literature suggests a convincing and unambiguous answer to the question raised in the title of this section: no, cities are *not* a thing of the past. It can even be asserted that cities are likely to remain one of the main engines of modern economic growth. Agglomeration economics are not disappearing but their nature and concrete form is changing. But even so, if agglomeration economies are that strong, at least in some sectors, why do cities have a finite size and why are there so many of them? As we are going to see, agglomeration economies have their dark side that restricts the process of city growth and leads to the emergence of a system of cities.

2.2 The trade-off between commuting and housing costs

In addition to the idea of agglomeration economies, two other fundamental concepts lie at the heart of urban economics: (i) *people prefer shorter trips to longer trips,* and (ii) *people prefer having more space than less space*. Since activities cannot be concentrated on the head of a pin, they are distributed across space. The first analysis of the way land is allocated across different activities was by von Thünen (1826), considered the founding father of spatial economics. The authoritative model of urban economics, which builds on von Thünen, is the monocentric city model developed by Alonso (1964), Mills (1967), and Muth (1969). Treading in these authors' footsteps, economists and regional scientists alike have developed the monocentric model in which a single and exogenously given central business district (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 city 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.

Commuting and housing are the two main consumption items for households in most developed countries. In the United States, the expenditure share on housing is 33 percent, while 18 percent is spent on transportation (Bureau of Labor Statistics). However, the housing share provided by the Bureau of Labor Statistics vastly exceeds the estimation made by Davis and Ortalo-Magné (2011) who find a 21 percent share, which is almost constant over the 1960 to 2006 period. Housing and transportation represent respectively 26 percent and 17 percent of French households' expenditures (INSEE). In Belgium, they account for 26 percent and 13 percent, respectively (Statistics Belgium). Admittedly, the expenditure share on transportation takes into account outlays unrelated to commuting, but it disregards consumers' time costs and the disutility associated with commuting.

Ever since the early 1970s, urban economics has advanced rapidly and has no sign of abating. The reason for this success is probably that the monocentric city model is based on a competitive land market. This assumption can be justified on the grounds that land in a small neighborhood in any location belonging to a continuous space is highly substitutable, thus making the competitive process for land very fierce. By allocating to some consumers a plot of land near the CBD, the commuting costs borne by other consumers are indirectly increased as they are forced to set up farther away. Hence, determining where consumers are located in the city is a general equilibrium problem. In equilibrium, identical consumers establish themselves within the city so as to equalize utility. In such a state, the land rent at a particular location is equal to the largest bid at that location. Since people are willing to pay more to be closer to their working place in order to save time and money on commuting costs, the urban land rent decreases with the distance from the CBD. In turn, since the land price is lower, the population density decreases with distance from the CBD because consumers can afford to buy more land. In sum, *the land rent reflects the differential in workers' accessibility to jobs*.

2.2.1 The monocentric city model

Consider a featureless plain with a dimensionless CBD located at $x = 0$ and a population of consumers who share the same income and the same preferences $U(z, s)$ where z is the consumption of a composite good, chosen as the numéraire, and s the amount of space used. Consumers compete to be as close as possible to the workplace, but the amount of land available in the vicinity of the CBD is too limited to accommodate the entire population. How, therefore, do consumers distribute themselves across locations? This is where the land market comes into play. The formal argument is disarmingly simple. Denoting by $R(x)$ the land rent prevailing at a distance x from the CBD and by $T(x)$ the commuting cost borne by a consumer residing at x, the budget constraint of this consumer is given by $z(x) + s(x)\, R(x) = I(x) \equiv Y - T(x)$, where consumers have by assumption the same income Y. Despite its simplicity, this model provides a set of results consistent with several of the prominent feature of cities.

Let $V(R(x), I(x))$ be the indirect utility of a consumer at x. Since the highest utility level attainable by consumers is invariant across locations, the derivative of $V(R(x), I(x))$ with respect to x must be equal to zero:

$$V_R R'(x) + V_I I'(x) = 0.$$

Using Roy's identity and the equality $I'(x) = -T'(x)$, we obtain the Alonso-Muth equilibrium condition:

$$s(x)R'(x) + T'(x) = 0. \qquad (1)$$

Since a longer commute generates a higher cost, this condition holds if and only if the land rent R(x) is downward sloping. As a consequence, (1) means that *a marginal increase in commuting costs associated with a longer trip, $T'(x)$, is exactly compensated for by the income share saved on land consumption*. In other words, people trade

bigger plots for higher commuting costs. If commuting costs were independent of the distance (*T'(x) = 0*), the land rent would be flat and constant across locations. In other words, *commuting costs are the cause and land rents the consequence*.

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 spatial equilibrium condition implies that the utility level is the same across all consumers. As a consequence, in equilibrium, the consumer optimization problem yields a compensated demand for land that depends on the land rent and the endogenous utility level. The utility level is treated as a given by every consumer who is too small to affect it. With housing a normal good, a lower price for land therefore implies higher land consumption. In other words, as the distance to the CBD increases, the lot size rises whereas the consumption of the composite good decreases.

The Alonso-Muth equilibrium condition (1) becomes especially revealing in the special case where consumers use the same exogenous lot size. Indeed, when *s(x)* is constant and normalized to one by choosing the appropriate unit of land, (1) can be rewritten as follows:

$$R'(x) + T'(x) = 0,$$

which implies that

$$R(x) + T(x) = \text{constant},$$

for all occupied locations *x*. In other words, the land rent is the mirror image of the commuting costs and capitalizes the advantage generated by a greater proximity to the CBD. When the lot size is variable, this relationship ceases to hold. However, the two magnitudes remain closely related at the aggregate level: If commuting costs are linear in distance, then the *aggregate differential land rent* is equal to *total commuting costs* when the city is linear, whereas it equals half the total commuting costs when the city is circular (Arnott, 1981). Thus, the former is positive if, and only if, the latter is positive.

Building on Alonso (1964), Fujita (1989) has developed an alternative approach that sheds further light on the working of land markets. A consumer's *bid rent b(x,u)* is defined by this consumer's willingness to pay for one unit of land at *x* while enjoying the utility level *u*. Formally, the function *b(x,u)* is defined as follows:

$$b(x,u) = \max_{z,s} \left[\frac{Y - T(x) - z}{s} \quad \text{s.t.} \quad U(z,s) = u \right] . \qquad (2)$$

Since *U(z,s)* is increasing in the consumption of each good, the equation *U(z,s) = u* has a single solution, denoted by *Z(s,u)*, which represents the consumption of the composite good when the land consumption is s. As a result, the bid rent function can be redefined as the following unconstrained maximization problem:

$$b(x,u) = \max_{s} \frac{Y - T(x) - Z(s,u)}{s} \tag{3}$$

whose solution gives the equilibrium land consumption $s^*(x,u)$ at location *x*. Other things being equal, when commuting costs at x decrease or the income increases (or both), the consumer offers a higher bid *b(x,u).* The bid rent also depends on consumers' preferences for land and the composite good through the expression *Z(s,u)*. Differentiating (3) at $s^*(x,u)$ with respect to *x* and using the envelope theorem yields

$$\frac{\partial b(x,u)}{\partial x} = -\frac{T'(x)}{s^*(x,u)} < 0. \tag{4}$$

Therefore, since the land rent is the upper envelope of the bid rent functions, the land rent decreases with the distance to the CBD. The land market thus works as if each consumer were to compare possible locations and evaluate, for each location, the maximum rent per unit of surface that this consumer is willing to pay to live there. Each plot is then occupied by the consumer offering the highest bid. This mechanism bears some strong resemblance to an auction as a location is a differentiated and unproduced good.

The monocentric city model also explains how the development of modern transportation methods (cars and mass transportation) has generated both suburbanization and a flattening of the urban population density, an evolution known as urban sprawl. The monocentric city model has thus produced results that are consistent with some of the main features of cities. The best synthesis of the results derived with the monocentric city model remains the landmark book of Fujita (1989). However, it is worth stressing that this model remains silent on why there is a district where all jobs are concentrated. So, we are left with the following question: *Do cities emerge as the outcome of a trade-off between agglomeration economies and commuting/housing costs?*

2.2.2 The emergence of employment centers

The first answer to this question was provided by Ogawa and Fujita (1980) in a fundamental paper that went unnoticed for a long period of time.[10] These authors combine consumers and firms in a full-fledged general equilibrium model in which goods, labor, and land markets are perfectly competitive. Informational spillovers act as an agglomeration force. Indeed, the value of a firm's location depends on its proximity to other firms because

[10] Only a limited number of papers have tackled the endogenous formation of employment centers. They are surveyed in Duranton and Puga (2004) and Behrens and Robert-Nicoud (2015).

informational spillovers are subject to distance-decay effect. As before, workers are keen to minimize commuting costs. 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 to work. In other words, the dispersion force stems from the interaction between the land and labor markets in firms' optimization program. 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.

Ogawa and Fujita show that, in equilibrium, *the city may display different configurations*. First, when commuting costs are high in relation to the distance-decay effect, the equilibrium involves a full integration of business and residential activities. To put it differently, land use is unspecialized and there is backyard capitalism. As commuting costs fall, two employment centers, which are themselves flanked by a residential area, are formed around an integrated section. Eventually, when commuting costs are low enough, the city becomes monocentric. In this configuration, land use is fully specialized. This seems to concur with the evolution in the spatial organization of cities observed since the beginning of the revolution in transportation. Activities were dispersed in pre-industrial cities when people moved on foot, whereas cities of the industrial era were often characterized by a large CBD. Modern cities retain a large CBD, but city centers now accommodate land-intensive activities performed in offices rather than factories that are big consumers of space.

Although the process of non-market interaction between firms (or workers) is typically bilateral, firms care only about their role as "receivers" and neglect their role as "transmitters." A comparison of the equilibrium and optimum densities shows that the former is less concentrated than the latter. This suggests that, from the social standpoint, the need to interact results in an insufficient concentration of activities around the city center. Therefore, contrary to general belief, firms and consumers would not be too densely packed. We note in passing that this result is in accordance with the literature on network economics where it is shown that the density of social networks is too sparse because building a link between two agents gives rise to an external effect that benefits the agents located in the vicinity of those involved in the new connection. Indeed, an agent who decides to build a link with another does not account for this effect, and thus socially desirable links may not arise (Jackson, 2008).

Since spillovers have the nature of a technological externality, *skilled jobs should be subsidized*. Quite the opposite happens: skilled jobs are overtaxed. Since national income taxes are based on nominal incomes, workers with the same real income pay higher taxes in large/high cost cities than in small/low cost cities because their nominal income is higher in the former than the latter. Unlike local tax differences, national tax differences are not compensated by a higher level of local spending, which tends to undersize the most productive cities. According

to Albouy (2009), the distortion generated by federal income taxes would lower long-run employment levels in high-wage areas by 13 percent, which is fairly substantial.

2.2.3 The urban system

Agglomeration economies explain why human activities are concentrated in cities. However, because commuting and housing costs rise with the population size, they—along with negative externalities generated by the concentration of people in small areas, such as congestion, pollution, noise, and crime—act as a strong force to put a brake on city growth. In accordance with the fundamental trade-off of spatial economics, *the size of cities may then be viewed as the balance between these systems of opposite forces*. Finding the right balance between the agglomeration economies and diseconomies is at the heart of the urban problem.

Not all cities are alike. The existence of very large cities in different parts of the world at different time periods is well documented (Bairoch, 1985). Cities have very different sizes and form an urban system that is *hierarchical* in nature: there are few large cities and many small cities, together with an intermediate number of medium-sized cities. The stability over decades or even centuries of the urban hierarchy is remarkable (Eaton and Eckstein, 1997; Davis and Weinstein, 2002). All cities provide private goods that are non-tradable (e.g., shops) and a variable range of public services (e.g., schools, day care centers). To a certain extent, the urban system reflects the administrative hierarchy of territorial entities. Because public services are subject to different degrees of increasing returns, cities accommodate a variable number of governmental departments and agencies, hospitals, universities, museums, and the like. More importantly, cities have a different industrial composition. In the past, cities produced a wide range of goods that were not traded because shipping them was expensive. Once transportation costs have decreased sufficiently, medium-sized and small cities got *specialized* in the production of one tradable good. This increased specialization often leads to significant labor productivity gains, but makes them vulnerable to asymmetric shocks. Today, only a few urban giants accommodate several, but not all, sectors.

Unlike specialized cities, *diversified* cities are better equipped when confronted with asymmetric shocks. Besides spillover effects between sectors, the coexistence of different sectors may also reduce the uncertainty associated with the initial phases of the product cycle (Duranton and Puga, 2001). For example, the preliminary stages in the development of a new technology or product require repeated contacts among those involved and such contacts are much easier when these people are in close proximity. Information becomes a spatial externality because, as it circulates within the local cluster of firms and workers, it inadvertently contributes to aggregate productivity. However, as shown by Helsley and Strange (2014), potentially beneficial clusters do not necessarily emerge, while the co-agglomeration that does occur in diversified cities may not be that which creates the greatest productive benefits.

In 1913, the German geographer 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, the second-largest city has on average about one-half the population of the largest city, a number 3 city one-third of that population, and a number *n* city has 1*/n* of that population. Formally, the rank-size rule holds that

$$\ln R_i = \alpha - \beta \ln P_i$$

where P_i is the population of city *i* and R_i its rank in the urban hierarchy. A large number of estimations of the coefficient α suggest a value very close to 1. Ever since Zipf (1949), it is now recognized that a *power* function provides a very good approximation of the population distribution (Gabaix and Ioannides, 2004). But is β = 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, which is still remarkable. The Zipf Law has attracted a lot of attention and what has been accomplished cannot be surveyed here (Gabaix and Ioannides, 2004).

In what follows, we briefly discuss a handful of contributions that aim to explain the existence of the urban hierarchy. 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 goods. However, this setup accounts for regularities only for special values of the structural parameters. The second-generation models explore the sorting of workers across cities as well as their composition, which are consistent with recent empirical evidence (Behrens et al., 2014; Eeckhout et al., 2014). However, the new models keep assuming that cities produce the same good or, equivalently, different goods that are traded at zero cost. Those various models 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.”

Thus, we find it fair to say that the dust is not settled down yet. However, we want to mention the work of Desmet and Rossi-Hansberg (2013) who decompose the determinants of the city size distribution into the following three components: efficiency, amenities, and frictions. Higher efficiency and more amenities lead to larger cities but generate greater frictions (congestion). This model may be used to simulate the effects of reducing variations in efficiency and amenities, which makes it is a relevant tool for designing regional and urban policies. Averaging the level of the above three components across cities and allowing the population to relocate leads to large population relocations but generates very low welfare gains in the U.S. Using the same model for China, the authors find much bigger welfare gains.

And indeed, recent research suggests that Chinese cities are “too small” (Au and Henderson, 2006). The *Hukou system*, which was established in China in 1951 and extended to rural areas in 1955, restricted rural-urban

migrations. In 1984, the central government allowed farmers to move to cities and to work in the manufacturing and service sectors. However, the system remains unchanged in nature because migrants must acquire different permits to access health care, schooling facilities and housing. Migrants are also imposed various hurdles to get those permits. Finally, they may still have to pay taxes to their home village for public services they do not consume. In other words, rural–urban migrants face total fees that can be equivalent up to several month wages. While most of these fees were abolished officially in 2001, various barriers still restrict labor mobility in China. As a result, migration costs remain high enough to support large wage inequalities.

The number of large metropolitan areas in the U.S. is proportionally much higher than in the EU. Therefore, it is tempting to follow *The Economist* (October 13, 2012), which argues that European cities are too small and/or too few for the European economy to benefit fully from the informational spillovers that lie at the heart of the knowledge-based economy. A more rigorous analysis has been developed by Schmidheiny and Südekum (2015). Using the new EC–OECD functional urban areas dataset, these authors show that, unlike the US urban system, the EU city distribution does not obey the Zipf Law. The reason for this discrepancy is that the largest European cities are "too small."

Undoubtedly, many European governments were not—and several of them are not yet—aware of the potential offered by their large metropolitan areas to boost national economic growth. Both in Europe and the U.S., "urbaphobia" has led governments to design policies deliberately detrimental to their large metropolises. In this respect, France is a good (or bad) case in point. For a few decades, Paris was considered as "too big" and public policies were designed to move away activities toward other French regions. European cities are a legacy of the past, which came into being on the occasion of the two great waves of urbanization, that is, the late Middle Age and the Industrial Revolution (Pirenne, 1925; Bairoch, 1985). By French standards, Paris is big. However, on the international marketplace, Paris competes with a great many comparable cities. Furthermore, it is not clear whether the formation of large metropolitan areas (10+ million people) is necessary to enhance European competitiveness. Much work remains to be done to understand the economic implications of the European urban space.

2.3 Urban policy: Some challenges

2.3.1 Housing

The choice of a residence implies a differential access to the various places visited by consumers. Therefore, it should be clear that the same principle applies when consumers are located close to locations endowed with amenities and/or providing public services such as schools and recreational facilities. As a consequence, if the general trend is a land rent that decreases as the distance from the CBD increases, the availability of amenities and public services at particular urban locations within the city affects this trend by generating rebounds in the

land rent profile (Fujita, 1989). For example, everything else being equal, if the quality of schools is uneven, the price of land is higher in the neighborhood of higher-quality schools. Likewise, dwellings situated close to metro stations are more expensive than those farther away.[11] All of this has a major implication: *in a city, 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*. This value is created by the community growth through actions taken by firms, households, and local governments, but not much value (if any) is created by the landlords.

As a first approximation, the value of a residential property may be viewed as the sum of two components: the value of the land on which the structure sits plus the value of the structure. The value of the residential structure has to belong to the agent responsible for its construction. In contrast, the land rent value depends on the proximity to jobs, as well as to public service providers, e.g., schools or the metro, which are financed by local or federal governments. Therefore, a laissez-faire policy that allows the landlord to capture the land rent is like an implicit transfer from the collectivity to the landlord. Evidently, for the land capitalization process to unfold, the land prices must be free to react to consumers' residential choices.

Assuming that consumers are identical, Stiglitz (1977) has shown that the land capitalization process is a very powerful instrument with which to finance the provision of public goods: the aggregate land rent equals the level of public expenditure if and only if the population size maximizes the utility level of the city's residents. Under these circumstances, public services can be financed by taxing the land rent. When there are too many consumers, this leads to higher land rents, generating a total land rent that exceeds the public expenditure. In contrast, when public expenditure exceeds the aggregate land rent, the population is below the optimal size.

The above result is known as the "Henry George theorem" based on the proposal in 1879 for a confiscatory tax on pure land rents by the American activist Henry George in his book *Progress and Poverty*.[12] When governments collect the increase in land rent that results from the provision of public services, they are able to offer a level of services such that the total willingness to pay for the services equals the marginal social cost. This promotes the efficient use of these services and reduces the distorting effects of taxes. From Léon Walras to the Nobel Laureates Franco Modigliani, Robert Solow, James Tobin, and William Vickrey, several prominent economists have argued that *land should be publicly owned*. Needless to say, such a radical recommendation is not politically acceptable.

---

[11] Using hedonic regression techniques, Diewert and Shimizu (2015) study housing prices in Tokyo and find that "for properties where the walk to the nearest subway station is 2–8 minutes, an increase in walking time of 1 minute decreases the land value of the property by 0.35%; for properties where the walk to the nearest subway station is 8–13 minutes, an increase in walking time of 1 minute decreases the land value of the property by 2.01%; and for properties where the walk to the nearest subway station is over 13 minutes, an increase in walking time of 1 minute decreases the land value of the property by 1.71%."

[12] For conciseness, we have chosen the simplest model of land capitalization. In more general settings, the result needs qualification. The reader is referred to Arnott (2004) and Fujita and Thisse (2013) for detailed discussions of the Henry George theorem.

Nevertheless, the design of sophisticated property tax schemes and the taxation of the surplus created by public investments could alleviate the burden of public deficits and reduce the distortions brought about by the land property rights that prevail in most European countries.

Most cities rely on various types of taxation that can be avoided by moving to the suburbs where tax rates are often lower. The result is that many core cities are in financial distress or depend strongly on federal government support. On this occasion, it is worth recalling that the gigantic transformation of Paris under the direction of George-Eugène Haussmann in the second half of the nineteenth century was financed by "the money … borrowed against future revenues that would result from the increased property values created by the planned improvements" (Barnett, 1986). What was possible then should be possible today, allowing our cities to finance, at least up to a certain threshold, the investments made to improve urban life.

Equally important, a better understanding of the land market allows one to shed light on an ongoing and heated debate in many European countries, namely, rent control. Contrary to a belief shared by the media and the public, the past and current rise in housing costs in many European 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. For example, the evidence collected by Glaeser and Gyourko (2003) in the U.S. suggests that "measures of zoning strictness are highly correlated with high prices," while Brueckner and Sridhar (2012) find large welfare losses for the building height restrictions in Indian cities. The beneficiaries of these restrictions are the owners of existing plots and buildings. 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.

Another bad (even, as some would say, dreadful) policy, which is very popular in some European countries, is the implementation of urban containment that hurts new residents by reducing their welfare level or that motivates a fraction of the city population to migrate away. In addition, by restricting population size, such policies prevent the most productive cities from fully exploiting their potential agglomeration effects. Again, the big winners of such land use regulations are the landlords who benefit from the imposition of an artificial rationing of land. 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, housing land in the South East of England was worth 430 times its value as farmland (Cheshire et al., 2014). Under such circumstances, the land rent level also reflects the "artificial scarcity" of land stemming from restrictive land use regulation. It is worth stressing here that, in many EU countries, the land made available for housing depends on municipal governments. Therefore, it is hardly a shock that decisions regarding land use vary with political parties (Solé-Ollé and Viladecans-Marsal, 2012).

High housing prices make the city less attractive. This may deter young entrepreneurs and skilled workers from settling there, which weakens the city's economic engine. Freezing rents—one of the most popular instruments used by political European decision-makers—renders the housing supply function more inelastic. Subsidizing tenants does not work either because the money transferred to the tenants tends to end up in the landlords' pocket when the elasticity of the housing supply is weak. *Providing affordable housing through the adoption of market-savvy land and construction policies is one of the keys to the future economic success of cities.*

Housing markets play a critical role in the workings of a city. The same holds for labor markets, which are often local in nature (Moretti, 2011). However, describing what has been accomplished in this field would take us too far from the main purposes of this survey.

2.3.2 Spatial segregation

Spatial segregation may be viewed as the sorting of individuals *within* cities, mediated by the land market. We have seen that the land market can be studied by using the bid rent function, which measures a consumer's willingness to pay for any particular location. This simple principle has far-reaching implications for studying the social stratification of cities.

(i) ***Rich versus poor consumers***. The argument developed above in the case of homogeneous consumers can be modified to show how the land market sorts out consumers by income. For simplicity, consider two income groups ($i = H, L$), the rich and the poor who share the same preferences $U(z,s)$, and the same commuting costs $T(x)$. Replacing Y by Y_i in (2) with $Y_H > Y_L$, we denote by $b_i (x, u_i)$ the bid rent function of the i-th income group, with $u_H > u_L$, and by $s_i^*(x,u_i)$ the associated land consumption. If $b_H (x, u_H) > b_L (x, u_L)$ at some x, the continuity of the land consumption within each group implies that $b_H (x + dx, u_H) > b_L (x + dx, u_L)$ holds. As a consequence, the income group i will occupy the city neighborhood where it outbids the other group. In other words, *consumers' income differences translate into spatial segregation*. The city is segmented into different neighborhoods in which consumers have similar characteristics, as envisioned by Tiebout (1956). This segmentation is the unintentional consequence of decisions made by a great number of consumers acting in a decentralized environment. This probably explains why many public policies that promote social mixing within cities fail to reach their objective.

However, we do not know yet how residents are sorted out. The social stratification of the city results from the ranking of the bid rent functions in terms of their slope in a sense that will now be made clear. Following the approach developed to obtain (4), we obtain

$$\frac{\partial b(x,u_i)}{\partial x} = -\frac{T'(x)}{s^*(x,u_i)}. \qquad (5)$$

At the boundary *x* between the two groups, the land rent must be the same for the two groups so that the two bid rent curves b_H and b_L intersect at *x*. Since land is a normal good, the land consumption of the rich exceeds that of the poor. It then follows from (5) that b_L is steeper than b_H at the crossing point. As a consequence, consumers of class *H* (*L*) outbid those of class *L* (*H*) on the right (left) side of *x*. To put it differently, other aspects being equal, *the rich choose a residence near the city limits and push the poor toward the city center*.

This ranking may come as surprise, for the poor live in the city area where the land rent takes on its highest values. Because they can offer higher bids, *the rich can always secure to themselves the locations they like best*, which implies that the poor are left with the remaining locations. The rich, because of their higher income, consume more land and more units of the composite good. By choosing a location near the CBD, they would face a high price for land that could reduce their consumption. Since their higher income compensates for their longer commute, the rich move away from the CBD where they enjoy both a bigger place and a higher consumption of the composite goods. This in turn implies that only the central locations remain available for the poor who consume small plots in the vicinity of the CBD. As this argument can readily be extended to any number of income classes, there is perfect *spatial sorting* of consumers within the city. Since competition on the land market is perfect while no externality is at work in the present setting, we may conclude that the spatial separation between the poor and the rich is not evidence per se of a market failure.

What is more, as shown by Glaeser et al. (2008), core cities in the U.S. attract the poor because they provide public transportation, while the core city governments tend to adopt policies more favorable to the poor than suburban governments. In other words, *core cities are poor because they attract poor people, not because they make people poor*. In sum, much as individuals are sorted across cities according to their skills, they are sorted across neighborhoods within those cities.

(ii) ***Heterogeneous commuting costs****.* In the foregoing, both groups of consumers bear the same commuting costs. However, rich consumers typically have a higher opportunity cost of time, so that the rich value accessibility to the workplace more than the poor. As a consequence, the trade-off between housing and commuting costs differs across income groups. For example, if commuting costs are given by $t_i\ T(x)$ where t_i is a positive constant, (5) becomes

$$\frac{\partial b(x,u_i)}{\partial x} = -\frac{T'(x)}{s^*(x,u_i)}. \qquad (6)$$

This expression shows that b_H is steeper than b_L at *x* if and only

$$\frac{t_H}{s^*(x,u_H)} > \frac{t_L}{s^*(x,u_L)}.$$

In this case, the residential pattern is reversed: the rich reside near the city center and the poor in the suburbs. Otherwise, we fall back on the previous pattern in which the rich choose to locate in the outer ring.

The above results show how the trade-off between housing and commuting works to shape the residential pattern among heterogeneous consumers. What is more, when there are several income classes with $Y_1 < Y_2 < Y_3$ and $t_1 < t_2 < t_3$, the income levels need not be perfectly correlated with the distance to the workplace. The market outcome now depends on how the demand for land varies with income and how the commuting parameters t_i differ across income classes. In other words, *incomplete sorting* may emerge as an equilibrium outcome, and thus there is social mixing across, not within, income groups.

The bid rent function has other major implications. Although some American core cities have rich enclaves, high-income residents in U.S. urban areas tend to live in the suburbs. This pattern is often reversed in the EU. Brueckner et al. (1999) have proposed an amenity-based theory that predicts a multiplicity of location patterns across cities. Europe's longer history provides an obvious reason why its core cities offer more amenities, such as buildings and monuments of historical significance, than do their U.S. counterparts. When the center has a strong amenity advantage over the suburbs, the bid rent function can be used to show that the rich are likely to live in central locations. When the center amenity advantage is weak or negative, the rich are likely to live in the suburbs. In other words, *superior amenities make the core city rich, while weak amenities make it poor*.

In the same vein, when the urban space is not featureless, the rich can afford to set up in locations with better amenities, which may be exogenous or endogenous, and with more transport options than the poor. In particular, decentralizing the supply of schooling may exacerbate initial differences between people by allowing the rich to afford better education for their children. This in turn tends to increase differences in human capital among young people and to worsen income inequality between individuals and neighborhoods within the same city. Besides income and preferences, spatial segregation as an equilibrium outcome can also be based on culture, race, and language. Through non-market interactions, the gathering of people sharing the same characteristics may generate different types of externalities, as in Schelling (1971). As in the foregoing, we end up with more homogeneous districts, but more heterogeneity between districts. When an education externality is at work, social segregation is the only stable equilibrium. It generates more income inequality, but may also give rise to efficiency losses for the rich and the poor through a misallocation of skills (Bénabou, 1993).

What makes spatial segregation a robust outcome is that, even in the absence of externalities, similar people competing on the land market will choose independently to be close to each other. The bid rent mechanism suggests that “causation runs from personal characteristics to income to the characteristics of the neighborhood in which people live” (Cheshire et al., 2014). Whether and how neighborhood effects have an impact on individual characteristics is an important topic, as European cities tend to become more polarized and segregated.

Surveying the on-going research would take us too far from the main purposes of the paper. Topa and Zenou (2015) review this field and stress the importance for policy of understanding the causality links and to distinguish between the neighborhood effects and the network effects. Neighborhood effects mean that a better accessibility to jobs reduces the unemployment prospects of the poor. This can be addressed by housing, transport or neighborhood regeneration policies. For example, distressed urban areas can be more or less isolated. This helps to explain why place-based policies, like the French enterprise zone programs, may increase the employment rate of the poor in well-connected areas, but not in rather isolated areas (Briant et al., 2015). Network effects have to do with the poor quality of the socio-economic group to which they belong. In this case, transport policy is useless and specific social integration and human capital policies are needed. Topa and Zenou (2015) point to empirical evidence for Sweden and Denmark, which suggests that ethnic enclaves can have positive effects on labor-market outcomes and the education level of immigrants, especially for the unskilled. The dark side is that such enclaves seem to have a positive impact on crime as growing up in a neighborhood with many criminals around has a long-term effect on the crime rate among immigrants.

To sum up, even though urban land use patterns reflect a wide range of possibilities, the way the bid rent functions varies with places' and residents' characteristics allows one to understand what kind of residential pattern emerges. The bid rent function, because it relies on a fundamental principle that guides consumers' spatial behavior, is likely to be useful in designing market-savvy policies fostering less segregation.

2.3.3 The organization of metropolitan areas

The spread of new cities in Europe came to an end long ago, so the European landscape has been dominated for a long period of time by a wide array of monocentric cities. European cities, probably because they were smaller than their American counterparts, undertook a structural transformation that is illustrated by the emergence of polycentric metropolitan areas (Anas et al., 1998). Indeed, the burden of high housing and commuting costs may be alleviated when secondary employment centers are created. Such a morphological change in the urban structure puts a brake on the re-dispersion process and allows big cities to maintain, at least to a large extent, their supremacy (Glaeser and Kahn, 2004). Among other things, this points to the existence of a trade-off between within-city commuting costs and between-cities transportation *costs*, which calls for a better coordination of transport policies at the city and interregional level.

Urban sprawl and the decentralization of jobs have given rise to metropolitan areas that include a large number of independent political jurisdictions providing local public goods to their residents and competing in tax levels to attract jobs and residents. A few facts documented by Brülhart et al. (2015) give an idea of the magnitude of this evolution. Metropolitan areas with more than 500,000 inhabitants are divided, on average, into 74 local jurisdictions, while local governments in the OECD raise about 13 percent of total tax revenue. Therefore, a cost-

benefit analysis of an urban agglomeration cannot focus only on the core city. Indeed, the metropolitan area is replete with different types of externalities arising from its economic and political fragmentation. As a consequence, what matters is what is going on in the metropolitan area as a whole.

The efficient development of a metropolitan area requires a good spatial match between those who benefit from the public goods supplied by the various jurisdictions and the taxpayers (Hochman et al., 1995). This is not often the case because a large fraction of commuters no longer live in the historical center. In other words, the administrative and economic boundaries of jurisdictions usually differ within metropolitan areas. Since constituencies are located inside the jurisdictions, local governments tend to disregard effects of economic policies that are felt beyond the political border, an issue that also arises at the international level. In addition to spillovers in the consumption of public goods, this discrepancy is at the origin of business-stealing effects generated by tax competition, which are studied in local public finance. However, this literature has put aside the spatial aspects that play a central role in the working of metropolitan areas. For example, the huge Tiebout-based literature leaves little space for urban considerations.

To the best of our knowledge, apart from a handful of papers discussed by Brülhart et al. (2015), urban economics is not used as a building block in models studying the workings of a metropolitan area. Thus research needs to be developed that recognizes the importance of the following aspects of the problem. First, agglomeration economies within core cities represent a large share of metro-wide agglomeration economies. This in turn implies that the CBD still dominates the metropolitan area's secondary business centers and attracts cross-commuters from the suburbs. As a consequence, agglomeration economies being internalized (even partly) in wages, the economy of the CBD generates some wealth effects that go beyond the core city and impact the suburban jurisdictions positively. Moreover, owing to the attractiveness of the CBD, the core city's government is incentivized to practice tax exporting through the taxation of non-resident workers. In doing so, the core city ends up with a CBD that is too small from the viewpoint of the metropolitan area. Therefore, the structure of the metropolitan area is inefficient as firms and jobs are too dispersed for the agglomeration economies to deliver their full potential (Gaigné et al., 2013).

Second, the suburbanites who work in the CBD benefit from public services provided in the core city but do not pay for them. This is a hot issue in cities like Berlin, Brussels, or Hamburg, which are also legal regional entities. Third, the metropolitan area is formed of local labor markets that are not well integrated and that coexist with pockets showing high and lasting unemployment rates. Fourth, and last, as cities grow, spatial segregation and income polarization tend to get worse. Whereas the social stratification of cities seems to be less of a political issue in the U.S., it ranks high on the agenda of many EU politicians and is a major concern for large segments of the European population.

The political fragmentation of metropolitan areas has other unsuspected consequences. The construction of large shopping malls and supermarkets in suburbia has exacerbated the extent of urban sprawl and contributed to the hollowing out of many city centers. In his classical work, *Downtown*, Fogelson (2005) writes that "the decentralization of the department store is one of the main reasons that the central business district, once the mecca for shoppers, does less than 5 percent of the retail trade of metropolitan areas everywhere but in New York, New Orleans, and San Francisco." That said, forces akin to agglomeration economies are at work in this context. Indeed, the entry of new firms into a marketplace generates a network effect that makes this place more attractive to consumers. However, this effect has more to share with the models of NEG discussed in Section 3 as it depends on market size. Equally important is the fact that consumers are attracted by marketplaces that are near because of travel costs. Combining the distance and network effects implies that consumers' shopping behavior is driven by gravitational forces.

Establishing new malls at the city outskirts diverts consumers from visiting downtown retailers. This in turn leads to a contraction of the central commercial district through the exit of retailers, which makes this shopping area even less attractive. The overall effect is to further reduce the number of customers, which cuts down the number of retailers once more. This vicious circle keeps on until no firm operates in the city center. In contrast, when high-income consumers (and tourists) spend more than low-income households on luxury and cultural goods and prefer to shop in small boutiques rather than in malls, modern amenities, such as restaurants and life performance theaters, choose to locate in the inner city. As local government can collect more taxes, they are able to provide better public services such as safe streets and good schools, thereby attracting more residents having higher levels of human capital. This type of virtuous circle helps shopping streets supplying high-quality goods and services to stay in business, whereas the vicious cycle associated with high crime rate, low-quality schools, bad transportation facilities, and the like will lead to the eventual disappearance of small shops (Ushchev et al., 2015). The fragmentation of metropolitan areas is likely to have accelerated the hollowing out of the city center.

2.4 Traffic and congestion

People travel within metropolitan areas for a wide range of reasons, such as commuting to work, dropping children off at schools, shopping in the CBD or suburban malls, and attending various family and social events. Even trade is much localized, thus implying a large flow of local shipments.[13]

The origin and destination of a trip, as well as the choice of a transportation mode, are decisions made by users. Economists study these decisions in a supply-and-demand context. The supply side is given by a transport infrastructure (roads, rail, airports), a transport service (bus, metro, taxi), and a price charged to the users (road

[13] For example, in the US the transport of goods within 5-digit zip code areas, which have a median radius of 4 miles, is 3 times larger than shipments outside the zip code area (Hillberry and Hummels, 2008).

user charge, parking fees, public transport prices). Users also supply personal inputs to their trips: cars, fuel, bicycles, insurance, and, most importantly, their own time. On the demand side, for every origin-destination pair, people travel for different reasons and have different opportunity costs of time. Since the supply of infrastructure is limited, the precise timing of trips also matters. It is, therefore, the total user cost of a trip (including money, time and discomfort) that ultimately determines an individual's demand for trips.

Most American cities (exceptions include New York; Washington, D.C.; and San Francisco) rely on car transportation, whereas public transport accounts for a significant fraction of trips in most European cities. This duality is reflected in the topics studied in the academic literature. In the U.S., where road pricing seems to be banned from public debate, there is more focus on the pricing of parking and optional varieties of road pricing like pay lanes. In the EU, even though some European cities have pioneered new congestion pricing schemes, national and local governments alike favor other policies such as high gasoline prices, as well as investments and subsidies in public transportation.

Urban transport issues can be studied from a short run or from a long run perspective. In the short run, the origins and destinations (residences, workplace, and shops) as well as the transport infrastructures (roads, rail, and subway) are exogenous, and thus policy options are restricted to pricing (fuel excises, parking, and rail tickets) and regulation (speed limits, pedestrian zones). Passengers can react via the number and timing of trips, as well as the type of transport mode. In the long run, locations are endogenous, as is the city size. By implication, users of the transport system have more options because they may change destinations (workplace, shopping) and origins (residence). The set of policy options is also much wider: one can add transport infrastructures and regulate the use of land (housing permits, type of activities). Most of transport economics focuses on the case where locations are given: how is the current infrastructure used (choice of mode, network equilibrium) and how can the policy maker improve the use of the existing infrastructures. There exist several types of externalities, and thus there is no satisfactory market mechanism to guarantee the best use of existing capacity. In addition, most road infrastructure can be accessed freely.

In what follows, we first consider the case in which locations and infrastructure are exogenous and focus mainly on passenger transportation. To be precise, we first define and discuss the estimation of external costs associated with transport trips for given origins and destinations and, then, look at public policies that can be used to address the various market failures associated with the supply and demand for trips. In the last two subsections, we discuss the policy issues when locations and transport infrastructure are endogenous. This will bring us back to the core question of urban economics: how to understand the organization of cities and the location of different activities.

2.4.1 External costs generated by the transport of urban passengers

Urban transportation accounts for some 20 percent of total passenger-kilometers, where a passenger-kilometer is defined as one passenger who carries one kilometer. In European cities, cars are the dominant transport mode (70 percent), while public transport (rail, metro, and bus) accounts for the remaining share. External urban transport costs are difficult to measure because they result from decisions made by a myriad of individuals who do not pay the full costs that their decisions impose on other users. One therefore has to rely on indirect measurements using connected markets (e.g., the variation of housing values as a function of traffic nuisances) or constructed markets (experiments and surveys). In the *Handbook of External Costs* published by the European Commission (2014), five types of external costs are considered: climate costs, environment costs, accident costs, congestion costs, and wear and tear of infrastructure. In Table 2, we document the relative importance of these costs for cars and public transportation (PT) in the EU.[14] Although the emission of greenhouse gases is proportional to the type and quantity of fossil fuels used, an open question remains about how to evaluate the damage generated by one ton of greenhouse gases, which is the same across industries, power generation, and the residential sector. In table 2, the climate damage generated by one vehicle kilometer is evaluated at €25/ton. In industry, the cap on greenhouse emissions has resulted in prices varying between 5 and 30 €/ton.

Table 2. External costs – orders of magnitude

External costs	**Costs in euro cents**	
	Cars (by passenger-kilometer)	**Public transport (by vehicle-kilometer)**
Climate cost	0.8	2.1 (bus)
Environment cost	4.3	21.4 (bus)
Accident cost	0.3	
Congestion cost	0.6 to 242.6	0 to 576.3 (bus)
Wear and tear infrastructure cost	0.8	2.7 (bus)

The external costs of air pollution are mostly health costs, while external noise costs also include disturbance and discomfort. For both categories, there is still an ongoing debate with epidemiologists regarding the magnitude of the health effects, and thus their effects as external costs (European Commission, 2014). The marginal external accident cost is also a matter of debate as it depends on two uncertain factors. First, to what extent are drivers

[14] This table gives orders of magnitude, for there are large variations within most types of external costs.

aware of their own expected physical accident risk and of their expected future insurance premiums in an experience-rated system? Second, what is the effect on the average accident rate of different types of accidents (car-car, car-bike, and the like) when an extra car is added to the road? A higher density of cars tends to make drivers more cautious. On the other hand, an extra truck increases the probability of a car-car accident because car drivers strive to avoid a collision with a truck.

Road congestion costs are the most important external costs generated in urban areas, but they also vary substantially across space and time. The marginal external cost generated by traffic congestion is the additional time, schedule delay, and resource costs borne by other road users when one additional user decides to travel by car. This type of external cost is poorly understood by the general public, probably because car drivers experience their own time loss. This time loss is internalized by the drivers, but the additional time loss incurred by the others is not taken into account by the individual drivers. In the simplest formulation, the average time cost of a road trip is given by $AC(X) = a + b\,X$, where X is the volume of traffic on a given road. If the total time cost is given by $TC(X) = X\,AC(X)$, the marginal social cost (*MSC*) is

$$MSC(X) = a + 2\,b\,X.$$

The marginal external cost is then equal to $MSC(X) - AC(X) = b\,X$. Since the road capacity is constant over the day, the marginal external cost is expected to vary a lot with the intensity of the traffic flow.

The wear and tear on the road is the damage caused by cars, which increases maintenance costs and discomfort for other drivers. Damage is mainly related to axle weight, so trucks and buses cause the majority of the damage.

Above all, Table 2 confirms *the sizable impact and variability of congestion costs compared with the other external costs*.

For PT, positive density economies arise when the frequency of service increases with demand. Higher frequency decreases the expected waiting time for passengers who arrive randomly at the bus stop and decreases the schedule delay time for non-randomly arriving passengers. PT by bus also contributes to congestion on the road. Because an additional passenger has to fit into the fixed capacity of the PT vehicles, there is also a negative discomfort external cost.

2.4.2 The difficult road to first-best pricing of congestion

First-best pricing means that all transport activities are priced (or subsidized) so that the marginal user cost equals the marginal resource cost plus the marginal external costs. As there are different types of external costs, this requires different types of instruments. The easiest external cost to internalize is damage to the climate because this cost is proportional to the consumption of fossil fuel. A fuel excise tax on gasoline and diesel is sufficient to

provide the right incentive to save fuel and, therefore, to reduce carbon emissions. That said, the most important marginal external cost of car use is congestion.

(i) ***Congestion*.** Ever since the pioneering works of Pigou (1920) and Vickrey (1969), economists have agreed that the *ideal instrument to tackle urban road congestion is congestion pricing*. The concept is easy to grasp. Many road transport externalities are strongly place-and time-dependent and, therefore, can hardly be tackled by using instruments such as fuel taxes or fixed car taxes, whereas congestion pricing is based on the on-going traffic. The first successful implementation of a congestion charge was in Singapore (1976). European cities that have introduced similar pricing instruments include London (2003), Stockholm (2007), Milan (2012), and Göteborg (2014). London has implemented zonal pricing; Stockholm, Milan, and Göteborg have implemented cordon tolls; but all systems have prices varying by the time of day. In a cordon pricing system, road users pay only when entering the zone, but trips within the zone are free. In a zonal pricing system, the user also pays for driving within a given zone. There have been heated debates in a large number of cities about adopting congestion pricing.

The application of road pricing is currently limited to only a few cities. So the question why implementing such a welfare-enhancing instrument fails is challenging. Of course, implementation of the pricing system and the transaction costs can eat away 10 to 20 percent of the toll revenues but technology is making big progress on this front. De Borger and Proost (2012) analyze the political economy of road pricing by means of a model of policy reform. Road users are unsure about the individual costs of switching from cars to PT. Three categories in the population are distinguished: the non-drivers; those who can easily switch to PT (the marginal drivers); and those drivers who have high substitution costs to switch to PT. Because non-drivers share the collected toll revenues, this population segment is always favorable to congestion pricing. Non-drivers and marginal drivers could form a majority for congestion pricing. However, ex ante, all drivers know only the average substitution costs to PT, which means that the marginal drivers expect a higher cost of switching to PT than what it will be. As a result, ex ante, there can be a majority of the population against congestion pricing. After implementation, however, the uncertainty is resolved. As a consequence, the marginal car users will see that their substitution costs are lower than what they expected ex ante, and thus may support congestion pricing ex post. Hence, a majority of drivers may vote against road pricing ex ante because their expected gain is negative, whereas a majority may support this policy after implementation.

Congestion pricing has been studied intensively in transport economics (Anas and Lindsey, 2012). Two lessons can be drawn. First, the design of the road pricing system is very important for the magnitude of the net welfare effect. For example, Stockholm was more efficient than London because the system had lower transaction costs and more finely differentiated charges over the time of the day. Indeed, time differentiation is crucial for capturing the full gains of congestion pricing. In the more detailed bottleneck model where homogeneous drivers trade off queuing costs and schedule delay costs by selecting a departure time, an appropriate toll scheme with strong time

differentiation can transform all queuing costs into revenue. The result would be an unchanged cost for the total trip and an unchanged total number of car trips, but departure times would be better distributed, and the local government would end up with extra tax revenues (Arnott et al., 1993). A simple differentiation of peak/off-peak times, as in London, would forego a large part of these gains and has to rely mainly on reducing the total number of peak trips to alleviate congestion. A more fine-tuned pricing scheme narrows the gap between the social benefits and the toll revenues. This is important for the political acceptability of peak pricing. For example, in London toll revenues may be a factor five higher than the net benefits, which generate strong lobbying against peak pricing or on how to share the collected toll revenues. More generally, *smart pricing of a bottleneck can transform queuing into toll revenue, bring about important time and productivity gains, and be a sensible alternative to the building of new and expensive transportation infrastructures*.

A second striking feature is that only a small proportion (25 percent or less) of the suppressed car trips was replaced by PT; the rest of the trips disappeared due to more car sharing, combining trips, or simply foregoing the trip (Eliasson et al., 2009). In the U.S., where urban congestion pricing is not implemented, high-occupancy lanes converted into HOToll lanes are spreading. In this case, one or more lanes of the highway can be used only by vehicles with two or more occupants and by those who pay for the use of the road. Having one or more of the lanes as toll lanes can be effective only if there is a sufficient difference in time values among users and requires a careful design of the tolls (Small and Verhoef, 2007).

First-best pricing of PT is comparatively easy to implement because every passenger has to enter a bus or metro and can be asked to pay. The resource costs and external costs of PT are complex but are known and vary strongly as a function of the density of demand and occupancy of the vehicle. For an almost empty bus, the cost of an additional passenger is limited to the additional time cost for the driver, the delay for the existing passengers and the other road users plus the climate damage. There is also a positive externality when additional passengers increase the frequency of the bus service and decrease the expected waiting time at the bus stop. In most urban areas, the largest external cost is probably the discomfort imposed by additional passengers in the peak period when the PT is close to capacity. First-best pricing would then require higher prices in the peak than in the off-peak time.

(ii) ***Parking*.** Besides traffic, *parking is another major source of urban congestion*. The supply of parking in a city takes up a lot of valuable urban land that could be used for housing and economic activities. A car is parked 95 per cent of the time and requires often a parking spot at the origin and at the destination. The parking supply is divided into parking available for everybody and the hypothecated parking.

On-street parking and commercial garages are usually available for everybody. When on-street parking is priced below the fare in parking garages, there will be search congestion and externalities that take the concrete form

of cruising for parking. First-best pricing of on-street parking implies on-street parking prices are comparable to (competitively) priced off-street parking. In this way, the negative externalities of cruising for parking disappear. One of the main changes over the last 20 years has been the privatization of enforcement of on-street parking. Enforcement became much more effective and the net revenues increased. New technologies allow the regular update of the prices of on-street parking. For example, in San Francisco, sensors keep track of the occupancy rate per block, which allows for the regular adjustment of the parking fees. The hypothecated parking is made available for residents, employees or shoppers. It is often provided for free, which worsens the unpriced congestion externalities. There have been many proposals to abolish these fringe benefits. A well-known example is the cash-out parking proposal where employers are forced to offer the option to receive the cash equivalent of free parking instead of free parking.

Supply of parking is often the result of second best minimum or maximum parking requirements that are decided by each city using some planning guidelines. Minimum parking rules are used to prevent developers and shopping malls to rely on unpriced residence parking nearby. Maximum parking regulations are used as second best policy to discourage car use. Guo and Ren (2014) found that abolishing the minimum residential parking requirement in London would lead to a reduction of residential parking supply of 40 percent. This is only an improvement if the nearby parking market is efficiently priced. In his review of the economics of parking, Inci (2015) finds parking one of these topics that is too important to be left to engineering and urban planners. As parking determines largely the role of cars in urban transportation (compare Los Angeles and New York), more research is needed on the effect of parking pricing and regulation in EU cities. Last, it is worth noting that parking rules also have unsuspected effects on urban forms (Brueckner and Franco, 2015).

External costs are defined for a given allocation of space among different transport modes, as well as for given regulations. For example, road space can be allocated among pedestrians (pavement), bike lanes, bus lanes, light rail, and cars. Regulations are about emissions per vehicle, speed limits, priority rules, and the like. Most economic research has focused on pricing, while most policy interventions focus on regulations and allocation of space. Optimizing transport flows requires the right combination of rules (say, speed limits), prices, and the allocation of space (e.g., bus lanes, on street parking).

2.4.3 The patchwork of policy instruments

In practice, we are far from first-best pricing schemes in urban transport. When it comes to transportation policies, the division of responsibilities among member countries, as well as regional and city authorities leads to a complex and knotty patchwork. The EU uses mainly regulation (car emissions, safety standards, and the like), while taxation power belongs to the member states. Cities have limited authority: parking fees, local traffic regulations, and subsidies for PT. Table 3 lists the most important taxes and regulations for road transport. PT is subsidized, sometimes very heavily, by member countries and cities, while fare structures are chosen by the PT companies

with some regulatory oversight by the subsidizing authority. As will be seen, many instruments are used but they are often poorly coordinated, not to say conflicting.

Table 3. The main instruments used for taxing and regulating road transport

Policy instruments	Cars	Trucks
Gasoline excise	Everywhere	
Diesel excise	Everywhere	
Tax and subsidies for other fuels	Lower tax (LPG) or no tax (electricity)	
Vehicle purchase and ownership taxes	Differentiated (size, environmental performance, type of fuel)	Differentiated (emissions, type of fuel, axle weight)
Parking charges	Differentiated (time and place)	
Distance charging		In some countries
Tolling	In some countries	In some countries
Road pricing by time of day and by place	In London, Stockholm, Milan, and Göteborg	In some cities
Regulation	Safety, emissions	Safety, emissions

The main tax instrument used to tax externalities of road use is the *fuel tax*. Even though this tax was probably established to raise public income (the average total revenue is 1.4 percent of the EU GDP), it is de facto the main tax instrument affecting the use of cars. If one considers the fuel tax on cars as the main instrument for correcting externalities, the tax should have the following second-best structure where the tax set is equal to all external costs associated with the consumption of a liter of fuel (Parry and Small, 2005):

$$\textit{Fuel tax/liter} = \textit{carbon damage/liter} + \gamma\ (\textit{kilometer/liter})\ (\textit{other external costs/kilometer}) \qquad (7)$$

The first term of this expression is the carbon damage that is proportional to the combustion of fossil fuel. When climate damage is assessed at €25/ton of CO_2, a low excise tax per liter (10 cents/liter) is sufficient. When there is no specific instrument used to price congestion and other externalities are related to distance driven rather than to fuel consumed, the only way to "price" these externalities is by adding an extra excise tax to the carbon tax for road use. This tax should equal the average other (non-climate) externalities related to road use, which explains the term (kilometer/liter) (other external costs/kilometer) in (7). To compute the tax per liter, one needs information on the other external cost per car kilometer and the number of kilometers per liter. Finally, one needs a correction factor *(γ)* that takes into account the share of fuel reduction due to reduced road traffic, not to more

efficient cars. Indeed, because congestion and accident externalities are related to distance rather than to fuel use, it is the amount of driving that the second component of the fuel tax aims to reduce, not the use of fuel itself.

To fix ideas, assume that an increase in the gasoline tax of 20 percent leads to a reduction in gasoline consumption of 10 percent of which 5 percent comes from a more fuel efficient car and 5 percent from less driving. Then, the factor *(γ)* equals 0.5. Assume, furthermore, that the other external costs per kilometer are on average 10 euro cents and that the car consumes 5 liters per 100 km. Under these circumstances, we obtain a second-best fuel tax equal to 10 + 0.5 x 20 x 10 = 110 euro cents per liter.

It is worth stressing that there is *an inherent conflict in using the gasoline tax to internalize both fuel-related externalities* (climate change) *and mileage-related externalities* (congestion, accidents). For climate-damage reasons, we want a car to be more fuel efficient (up to a marginal cost of €25/ton of CO_2). But to make car drivers take into account the other externalities, we want them to keep paying the same tax per kilometer. Using the gasoline tax to internalize two externalities at the same time leads to cars that are too fuel efficient while the other externalities are not fully internalized. Gasoline is taxed at 200 percent or more (1 Euro tax/liter on top of a production cost of 0.5 Euro/liter), which is equivalent to a tax of €300/ton of CO_2 (to be compared with €25/ton of CO_2, as recommended in other sectors of the economy). It looks as if we are at the limit of what can be achieved with a single second-best instrument. This should not come as a surprise: ever since the work of Tinbergen, it is well known that relying on a single instrument to pursue different objectives is likely to lead to inefficient outcomes. Given that the main objective of the gasoline tax is probably to collect tax revenue, using this tax as an instrument to solve all these problems amounts to squaring the circle.

It is not only the pricing of gasoline that went wrong; the pricing of diesel fuel for cars is also problematic as low diesel excise taxes led to the massive introduction of diesel cars in most of Europe. Diesel cars have a small carbon emission advantage but are more dangerous in terms of health damage when one relies on the real world emission results rather than on the results of the test cycle (ICCT, 2013). The U.S. took another route and has almost no diesel cars.

One of the most effective additional instruments to control the environmental externalities of car use is the regulation of emissions of traditional air pollutants. The Auto-Oil program of the EU regulated the emissions of new cars and the quality of fuel. This was efficient to tackle traditional pollutants (N0x, SO_2, particulates). By installing additional equipment (catalytic converter, lower sulfur content of fuels) at relatively low cost, emissions could be reduced by a factor of 5 to 20 (Proost and Van Dender, 2012). As for gasoline, the EU could benefit from the American and Japanese experience and technologies.

A complement to stricter air pollution regulations is the use of low-emission zones. In a low-emission zone, only the cleanest cars are allowed to move freely, while dirtier cars have to pay a charge or, if they get caught, a fine. As air pollution damage is directly proportional to the population density, it makes sense to have an additional instrument for dense urban areas. The EU ambient air quality regulation sets a maximum for the concentration of air pollutants and, when this maximum is exceeded, city or national governments have to take action. More than 50 German cities have experimented with different policy measures. The overall conclusion was that improvements in public transport were not effective, but access restrictions for dirty cars were (Wolff, 2015). This type of instrument is at present less effective because, over time, all cars will comply with the latest EU emission standards. But, as attention to conventional air pollution in cities is increasing and as the marginal cost of greening cars is increasing, this instrument could again become more useful. It allows for the differentiation of requirements for urban road traffic and non-urban road traffic. Instead, one could think of banning diesel cars and even gasoline cars in dense cities.

Using fuel-efficiency regulation for cars to reduce greenhouse gas emissions is a costly type of regulation as transport has already a high carbon tax under the form of the gasoline tax. One possible justification is the possible myopia or fuel efficiency gap. If consumers underestimate systematically the future fuel costs, a fuel efficiency regulation would help the consumers and would better signal the external costs. But the empirical evidence for consumer myopia is very weak for the EU car buyers. Grigolon et al (2014) analysed car buyer behaviour in the EU and found that consumers take 90 percent of the future fuel costs into account when they select a car. When this is combined with a fuel tax that is more related to the mileage externalities than to the fuel related externalities, imposing more fuel efficient cars is not an efficient policy measure. The EU is a world leader in terms of fuel-efficiency standards. If the aim is also to successfully transfer technology, we may need to re-orient our technology standards toward less ambitious targets because other countries have less ambitious climate objectives and do not want to pay for elaborate super-efficient technologies (Eliasson and Proost, 2015). Note also that many countries have used vehicle purchase and ownership taxes as additional instruments to reduce CO_2 emissions. The Netherlands, Denmark, Sweden, and France used vigorous policies to achieve significant carbon emission reductions but there is evidence that these policies were very costly and not effective.[15]

2.4.4 Public transport pricing

In the EU, PT accounts for a significant share of commuters. In most European cities, the recovery of operational costs is low (below 50 percent), while the peak demand is close to the rail and metro capacity. Implementing low prices for PT in cities is often presented as a good illustration of second-best pricing. But is such a recommendation

[15] For more details, see D'Haultfoeuille et al. (2013) and Munk-Nielsen (2014).

well grounded? In the expression (8), the optimal PT price $P_{PT,peak}$ is equal to the social marginal cost $MC_{PT,peak}$ of PT, corrected by the pricing gap between the price $P_{\text{car},peak}$ and the social marginal cost $MC_{car,peak}$ of car use. Computing the social marginal cost of a PT trip is not simple. Indeed, it requires to take on board scale economies (using available seats in metro or bus) and negative discomfort economies when vehicles are crowded. It must also account for the following positive economies: even when busses or metros are not full, it is optimal to raise frequency as this allows one to reduce waiting time (Mohring, 1972). In the absence of congestion pricing, the price of car use in the peak period is lower than its social marginal cost, so a subsidy for PT is efficient insofar as this subsidy is able to make car users switch to PT. For this, we need the fraction *f* of new PT users who would, in the absence of the subsidy, be car users:

$$P_{PT,peak} = MC_{PT,peak} + f \cdot (P_{car,peak} - MC_{car,peak}). \quad (8)$$

Parry and Small (2009) have found that a subsidy of close to 90 percent of the average operational costs for urban rail transport is socially desirable when *f* = 0.5, which seems to ground the proposal of strongly subsidized PT. These authors find that the subsidy is efficient for two reasons. First, there are important scale economies, which are the most important element to justify subsidies in the off-peak period. Second there are important unpriced car congestion externalities, which are the main reason to justify subsidies in the peak period.

However, some empirical studies find values for *f* that are smaller than 0.2 (van Goeverden et al., 2006). In this case, the optimal subsidy for the peak falls from 90 to 10 percent, thus casting serious doubt on the relevance of subsidizing the use of PT. In a numerical study for London as well as Santiago de Chile, Basso and Silva (2014) compare the pricing of car and bus combined with other instruments (bus subsidies, dedicated bus lanes, and congestion pricing). They find that dedicated bus lanes can be a much more efficient instrument than PT subsidies and are, in terms of efficiency almost as efficient as road pricing for Santiago de Chile. Results tend to be city specific as they depend on the current modal shares and the ease of substitution.

Current marginal fares for PT in the EU are often zero as most users pay a monthly subscription price, which allows them to travel when and as much as they want giving rise to massive congestion problems in PT systems of big EU cities (London, Paris).There is a need to look for more efficient pricing systems that account for the differences in cost between peak and off-peak trips and in function of area and distance traveled, as well as in function of the congestion levels of car transport. This could alleviate the financial problems of urban PT organizations Bus systems exhibit smaller scale economies so that correct peak load pricing can make them break even more easily. Rail and metro systems have increasing returns to scale. Hence, first-best pricing leaves them with a higher 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 financing gap that characterizes almost all PT

systems. As long as attention is payed to who pays for the subsidies to PT, there is not necessarily a conflict between more correct PT fares and redistribution policies (Mayeres and Proost, 2001).

In the last 20 years, the United Kingdom has experimented with privatized PT services. In London, bus services were tendered to private companies but one central bus authority remained to decide on schedules and prices. The end result was an important reduction in costs. Outside of London, bus services were fully privatized with the private companies deciding the entry, scheduling, and prices. There are only limited and targeted subsidies. As each bus company offers services at different times of the day, there is a clear tendency to offer higher frequency. By offering a time schedule that closely matches the timetable of a competitor, one company could steal passengers from other companies, but this did not turn out to be in the interest of the passengers. The end result was lower costs, higher prices, higher frequencies, and less competition (Mackie et al., 1995). For the London Underground, a public private partnership was chosen. Two firms were awarded contracts for maintenance and infrastructure while the public London Underground company remains responsible for train operations. This proved to be a difficult contract. Contracting out the operation of buses is more common than for rail and has led to important efficiency gains when the contracts are well designed (Gagnepain et al., 2013).

One may wonder if it makes sense to tax congestion while having high labor taxes. Parry and Bento (2002) find that charging the full external congestion cost to commuters remains the best policy as long as the additional tax revenues are used to reduce the existing labor tax. This type of reform makes commuters choose the right mode and volume of transport when they can choose between congested roads and a non-congested alternative mode of travel priced at marginal cost. This reasoning holds when only commuters travel in the peak and when there is a perfect substitute that is uncongested, such as PT (Van Dender, 2003).Furthermore, in the presence of agglomeration economies, the optimal congestion toll should be lower than the marginal external congestion cost because it tends to reduces employment in the city center, unless other instruments (subsidies to firms) are used to correct the agglomeration externalities. If fine-tuned road pricing implies only small shifts in working hours, then agglomeration externalities are not really affected (Arnott, 2007).

Finally, in many EU countries, a company car is proposed by employers as an untaxed fringe benefit, which leads to excessive car use and some employers also pay for all public transport expenses of their employees (Harding, 2014). All of this shows *the need of a global assessment of commuting expenses in relation to income tax*.

2.4.5 Does building new infrastructure reduce congestion?

To the public and many decision-makers, the answer seems obvious and positive. However, things are not that simple. First, when origins and destinations are given, more capacity leads to more car users. Hence, the time benefit of road extensions in the presence of unpriced congestion is reduced by this induced demand (Cervero, 2003). This already suggests that the standard approach to controlling congestion—forecast traffic growth and

build enough road capacity to accommodate it—is likely to be ineffective. Second, Arnott (1979) shows that improving transportation in a congested monocentric city leads to a new residential equilibrium in which congestion at *each* location increases compared with the initial equilibrium. In other words, once it is recognized that consumers respond to changes in commuting costs, building new transportation links loses a great deal of its appeal.

Duranton and Turner (2011) observe that those who argue in favor of new transportation infrastructure forget the simultaneity problem that we have encountered in studying agglomeration economies: the supply of roads and the density of traffic are interdependent phenomena. When the number of vehicles on the road is given, additional capacity decreases the density of traffic and makes trips faster. However, a higher capacity attracts more traffic, and thus density increases. All this implies that it is a priori unclear how the causality runs. This has led Duranton and Turner to study the congestion problem in American cities for the years 1983, 1993, and 2003, using modern econometric techniques. Their conclusions cast serious doubt on the merits of infrastructure-based congestion policies. First, Duranton and Turner confirm that new roads and public transit generate more traffic. What is less expected, but more important, is that, in the absence of road pricing and for some types of roads, "new road capacity is met with a proportional increase in driving." But where do the additional travelers come from? Again, the answer is not the one that comes immediately to mind: "the law of traffic congestion reflects traffic creation rather than traffic diversion." New cars and new trucks share the responsibility for the extra trips almost equally. Last, whenever the road capacity is extended and road use is not appropriately priced, the road extension will attract PT passengers. This reduces frequency in PT, a vicious circle that may lead to the disappearance of the PT alternative (Arnott and Small, 1994).

In sum, the works by Arnott, Duranton, Turner, and others have a major implication that runs against standard policy recommendations: when road pricing is not implemented, *building new roads need not be the appropriate policy to reduce traffic congestion*. Therefore, congestion pricing is back to center stage as the main tool to curb urban congestion. Despite the lack of enthusiasm of public policy-makers for this instrument, the large number of results obtained by urban transportation economics should encourage governments and other authorities to evaluate new transportation projects against smart pricing schemes.

Whenever one considers extending current road or PT infrastructure, one should keep in mind that new technologies may enhance the effective capacity of the existing transport system (Winston and Mannering, 2014). For example, the capacity of current road infrastructure may be enhanced by software applications that facilitate ridesharing. In the long run, vehicle to vehicle communication may increase the capacity of a road network by coordinating conflicting traffic flows and by using the stock of cars more intensively, freeing urban space from parking. Finally, the external air pollution and climate costs of road transport can decrease by switching to electric vehicles. For short distances electric bikes already offer a clean, cheap and fast (25 to 40 km/h) solution. In the

case of public transport, new technologies may also lead to a better use of existing capacity. In the short term, better software may generate "on demand" collective transport. Whenever there is a capacity shortage, pricing is crucial to use capacity optimally but road pricing also stimulates the development of the new technologies.

2.4.6 The wider benefits of urban transport projects and new developments in assessment methods

There is growing empirical evidence that big urban transport projects lead to changes in the city form. 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 percent decline in central city population between 1960 and 2011. In addition, a highway ray fostered a 20 percent population growth in the suburban municipalities where ramps were located. Finally, each additional kilometer close to the nearest highway ramp increased municipal density growth by 8 percent. This provides strong evidence for the role of highway capacity on the population distribution within the urban area.

It is, therefore, important to understand the full impact of a large transport project (or important traffic regulation) on the welfare of the metropolitan population, including efficiency as well as equity aspects. Planners have typically little faith in the efficiency or equity of market-determined outcomes, and advocate detailed land use planning. Yet, as argued in the urban economics section, *market forces drive land use to its most productive use if markets are corrected for the most important externalities*. However, care is needed in selecting which externalities to correct. For example, compact cities are often advocated to reduce carbon emissions generated by private transport. However, in the EU we have seen that carbon is already over-priced via the gasoline tax (see 2.4.3). What is more, in 30 years from now, standard cars might well be electric battery cars. So, climate considerations are not a good motivation for compact cities.

Economists have developed cost-benefit analysis (CBA) techniques that aim to assess transport projects, be it new infrastructures, new pricing or new regulations. In the EU member countries, they are now widely used, but not necessarily followed by the decision makers. The CBA techniques have progressed over the last 50 years from the Dupuit consumer surplus measures to methods that correct for externalities, as well as for market imperfections and the opportunity cost of public funds. Quinet and Raj (2015) review the progresses made in assessment methods and distinguish between three approaches: (i) the basic CBA method focusing on the changes on the transport market, corrected for externalities and side effects on other markets; (ii) the econometric analysis of the causality effects; and (iii) a detailed spatial modelling embedded in land use planning models. For non-marginal projects, such as large transport network extensions, there is a need to use them all.

Land use planning models have been around for a long time (Lowry, 1964). However, there is a need for operational models integrating both land use and transport (LUTI). Indeed, new transport infrastructures often

increase the demand for land, while there is often a new demand for infrastructures when new land is made available for urban activities. 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. Finally, it would be futile to seek a model based on a unified theory of cities that would appeal equally to economists, geographers, architects, and urban planners (Batty, 2013). 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 (Anas and Kim, 1996; Anas and Liu, 2007; de Palma et al., 2015).

In principle, LUTI models help to understand the effect of one particular policy intervention and ultimately answer the important question of the ideal urban form. 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 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. Moreover, most analyses focus on an ideal government planner while, in the real world, the political authority is dissipated over sometimes overlapping jurisdictions.

The new LUTI model developed by the CPB in the Netherlands provides a nice example of what can be accomplished in terms of a detailed understanding of the effects associated with a given policy. Teugels et al. (2014) built and estimated a LUTI model using data on transport infrastructure, commuting behavior, wages, land use and land rents for 3000 ZIP-codes in the Netherlands and for three levels of education. The city of Amsterdam and the Schiphol airport area are both located south of a major canal that connects the Amsterdam harbor to the North Sea. Today, the main connections between Amsterdam and the area north of the canal consist of five highway tunnels and two train tunnels. The above model is used to analyze the ex post impact of adding two more train tunnels. Differently from transport models that limit the analysis to a modal shift from road to train by commuters, the LUTI model accounts for the relocation of firms and households. Adding two train tunnels would lead to a higher concentration of jobs in Amsterdam and make the northern area more attractive to residents. The relocation effects appear to have a strong impact on total welfare as they add 40% to the modal shift benefits. Furthermore, as the high-skilled workers are more mobile and prefer to travel by train, they will benefit more from the new rail tunnels than the low-skilled workers.

2.5 Where do we stand?

Cities—but not all of them—have been and still are the main engines of cultural, economic, and social development. By encouraging social interactions and the exchange of ideas, cities allow for a finer division of labor and the quick adoption of innovations. As new ideas are often a new combination of old ideas, connecting people remains crucial for the Schumpeterian process of innovation to unfold. As human capital is the main production factor in knowledge-based economies, ignoring the role played by cities often leads governments to design policies that are harmful (but not on purpose!) to the economic fabric of their countries.

Not all cities are equally affected by innovation and growth; inequality cuts through the urban system. If anything else, the development of human capital should be the main target of urban policies. As accurately argued by Glaeser (2011), the oversupply of structures and infrastructures is the hallmark of stagnating and declining cities. Rather than spending billions of euros on large infrastructures and fancy buildings, local governments should facilitate movement in cities by means of congestion pricing and promote the supply of affordable housing. What is more, housing and transportation markets are intimately intertwined with local labor markets (Zenou, 2009; Ioannides, 2012). Therefore, European and national employment policies that ignore the urban environment in which jobs are created are likely to be unable to deliver their full potential. Moreover, understanding how the process of land capitalization works might help finance local public goods and services, thus alleviating the need to reduce city budgets because of macroeconomic fiscal constraints. In a nutshell, as Cheshire et al. (2014) write, "urban policy informed by economic insights can help improve policymaking for individual cities and urban systems as a whole."

All regions benefit from the agglomeration effects arising in large cities through interregional and interpersonal transfers. For example, in 2012, the Ile-de-France (Paris) produced 30.1 percent of the French GDP but received only 22 percent of the disposable income. In other words, 8 percent of the GDP is redistributed toward other French regions. Greater London's share of the GDP in the United Kingdom is 23.1 percent while its share in the U.K.'s disposable income is about 16.7 percent. In Belgium, the contrast is even more striking. The NUTS-2 region Brussels-Capital produces 20.6 percent of the Belgian GDP but receives only 10.3 percent of the disposable income; thus, more than 10 percent is redistributed toward the other two regions of Belgium. Very much like some American cities, Brussels attracts high-income commuters as well as poor residents.

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. Regional policies fail to recognize that regional income differences are often the result of scale economies. To a certain extent, this explains the disillusion regarding the effectiveness of policies that aim for a more balanced distribution of activities across the EU.

Urban policies are not within the competences of the European Commission. In contrast, regional policies are. While things are different when one moves from the microscopic to the macroscopic, we will see in the next section that they are not *that* different.

2.6 The need for more and better urban data in the EU

Although we recognize with Cheshire and Magrini (2009) that "it is inappropriate to argue that there is one unified European urban system", there is a need for more scientifically grounded empirical works on cities at the EU level. This requires the availability of several types of data. First, for comparative studies across cities to be meaningful, member countries should agree on the same geographical definition of what a metropolitan area is, as in the U.S. where the concept of "statistical metropolitan area" is widely used. Paul Cheshire's work on "functional urban regions" should be a source of inspiration to others. Similarly, local data about employment, transport, GDP, human capital, physical attributes (buildings, roads), environmental quality (air quality, soil), and cultural amenities should be made available for more countries. European economists very often study American cities rather than European cities because very good data are available in the U.S., but not in the EU. There is also a need for data at a fine spatial scale about what is going on within cities. For example, such data are needed to study how firms and households choose their locations. New technologies of data collection can help to overcome the data gaps and definitional problems in Europe.

3. Are regional disparities a bad equilibrium outcome?

The EU has a wide diversity of cultures and a wide range of incomes at the interregional level. Cultural diversity is an asset that has its costs and benefits, but sizable income differences are a source of concern. Article 158 of the Treaty on European Union states: "*the Community shall aim at reducing disparities between the levels of development of the various regions and the backwardness of the least favoured regions or islands, including rural areas."* European integration is supposed to lead, through more intense trade links, to convergence of income levels across countries. However, this process is slow and may be accompanied with widening interregional income gaps despite EU regional policy efforts.[16] The lack of regional convergence may lead to cohesion problems that, when combined with cultural differences can contribute to secessionist tendencies and threaten the future both of countries and of their membership in the EU. Whether or not there is convergence across the European regional system remains a controversial issue that also raises various unsuspected methodological difficulties (Magrini, 2004).

[16] See Boldrin and Canova (2001), Midelfart-Knarvik and Overman (2002), and Puga (2002) for early critical assessments of the EU regional policies.

If urban policies have not been greatly influenced by economic theory, there has never been a shortage of economists' speculating whether there is too much spatial inequality at the interregional level. It is ironic that, for a long time, regional concepts, models, and techniques were mere extensions of those used at the national level, with an additional index identifying the different regions (e.g., interregional input-output matrices). What is more, the fact that economists have emphasized commodity trade rather than production factor mobility and, despite Ohlin's recommendations, have neglected trade obstacles such as transportation costs has been an impediment to the development of regional economics. As a consequence, it has lagged way behind urban economics in terms of scientific content. Today, thanks to the appearance of NEG, we are better equipped to understand the uneven development of regions.

The idea of spatial interaction is central to regional economics. Broadly defined, spatial interaction refers to a wide array of flows subject to various types of spatial frictions. Examples of these flows include traded goods, migration, capital, interregional grants, remittances, as well as the interregional transmission of knowledge and business cycle effects. The bulk of NEG has been restricted to the movement of goods and production factors. NEG remains in the tradition of trade theory as it focuses on exchanges between regions to explain why some regions fare better than others. Furthermore, NEG models regions as dimensionless economies without land. In contrast, an approach that would build on urban economics would rather choose to focus on the internal functioning of a region. Both approaches are legitimate and a full-fledged model of the regional system should take them both into account (Storper, 2013).

The economic performance of regions is affected not only by their industrial mix and their relative position in the web of relations, but also by the interregional and international mobility of commodities and production factors (e.g., capital and labor). In particular, lowering transport and trade costs changes the incentives for both firms and workers to stay put or move to another location. Therefore, to assess the full impact of market integration and of the monetary union, it is crucial to have a good understanding of how firms and workers react to lower trade and transport costs. In this respect, it should be stressed that European policy-makers often overlook the fact that market integration affects the locational choices of firms and households. In particular, as will be seen, NEG highlights the fact that a rising mobility of goods and people does not necessarily reduce spatial inequality. Even though regional development agencies typically think of spatial inequality as "temporary disequilibrium" within the economy, stable spatial equilibria often display sizable and lasting differences in income and employment, a fact that agrees with casual evidence. Furthermore, we will see that regional disparities need not be bad because they can be the geographical counterpart of greater efficiency and stronger growth.

At the interregional and international scales, accessibility to spatially dispersed markets drives the location of firms; this has long been recognized in both spatial economics and regional science (Fujita and Thisse, 2013). Accessibility is itself measured by all the costs generated by the various types of spatial frictions that economic

agents face in the exchange process. In the case of goods and services, these frictions are called *trade costs*.[17] Spulber (2007) refers to them as "the four Ts": (i) *transaction costs* that result from doing business at a distance due to differences in customs, business practices, as well as political and legal climates; (ii) *tariff and non-tariff costs* such as different anti-pollution standards, anti-dumping practices, and the massive number of regulations that still restrict trade; (iii) *transport costs* per se because goods have to reach their destination, while many services remain non-tradable; and (iv) *time costs* because, despite the Internet and video-conferences, there are still communication barriers across dispersed distribution and manufacturing facilities that slow down reactions to changes in market conditions. Because they stand for the cost of coordinating and connecting transactions between the supplier's and customer's locations, trade costs are crucial to the global firm and therefore likely to stay at center stage. The relative importance of the "four Ts" obviously varies enormously from one sector to another, from one activity to another, from one commodity to another.

Anderson and van Wincoop (2003) provide a detailed estimate of trade costs, concluding that these costs would climb to approximately 170 percent of the average mill price of manufactured goods, but the variance across goods is high. This estimate can be broken down as follows: 55 percent internal costs, which include all logistics costs; and 74 percent international costs (1.7 = 1.55 × 1.74 - 1). International costs in turn are broken down into 21 percent for transport costs and 44 percent for costs connected with border effects (1.74 = 1.21 × 1.44). Tariff and non-tariff barriers account for 8 percent of the border effects (exceptionally, this is 10 or 20 percent in the case of developing countries); language differences, 7 percent; currency differences, 14 percent; and other costs, including information, 9 percent (all in all, 1.44 = 1.08 × 1.07 × 1.14 × 1.09). It is not exaggerating, therefore, to say that the share of trade costs in the consumer price of several manufactured goods remains high. Note that there are also big differences from one trading area to another. For example, Head and Mayer (2004) argue convincingly that North American integration is significantly deeper than European integration.

3.1 Interregional trade and transport

Transport, by its very nature, is linked to trade. And, as trade is one of the oldest human activities, the transport of commodities is therefore a fundamental ingredient of any society. People get involved in trade because they want to consume goods that are not produced nearby. The Silk Road is evidence that shipping highly valued goods over long distances has been carried out for this precise reason. The huge development in trade preceding World War I has led many economic historians to underline the emergence of a first phase of globalization in the second half of the 19th century, ending in 1914, largely explained by the dramatic drop in transport costs (O'Rourke and

[17] We follow the literature and view market integration as a gradual reduction in the costs of shipping goods and services.

Williamson, 1999). It is therefore legitimate to ask how transportation economics highlights our understanding of trade and vice versa.

3.1.1 The gravity equation

In an unnoticed chapter of in his classical book, Cournot (1838) proposed a simple trade model with one homogeneous good (and a numéraire) and several regions (or countries) that each correspond to a node of a transportation network. In each region, firms' and consumers' behavior is aggregated into demand and supply functions, while shipping the good from its origin to a destination is costly. Introducing transport costs into trade models seems simple and natural, natural because shipping commodities across space requires resources, and simple because transport costs could just be one more type of cost to take into account. But, as encountered by Samuelson (1954) himself, this feat appears very difficult because transport costs are associated with general equilibrium effects across spatially separated markets. Trade being driven here by spatial arbitrage, in equilibrium the price of a good at one place depends on the price for the same good in another location as arbitrage limits the price difference to the shipping cost of the good.

The competitive equilibrium is reached when the demand price in the importing region equals the supply price in the exporting region plus the unit transportation cost from the latter to the former. If the demand price is less than the supply price plus transportation costs, then no trade occurs. Evidently, when transportation costs are high, each region operates under autarky. Once these costs have decreased sufficiently, trade across regions comes into play. As the integration process deepens, some regions stop producing the good to become importers because the domestic producers are less efficient than their foreign competitors. In other words, these firms are driven out of business because they have lost "the most effective protection of all tariff protections, namely, that provided by bad roads" (Launhardt, 1885). By freeing resources for other production activities, technological progress in transport allows an increase in the production of consumption and intermediate goods. However, decreasing transport costs also redraw the production map, with some regions producing more and others less. This simple model highlights the importance of the three major forces stressed in modern trade literature: (i) the *size* of regions, through their demand schedules; (ii) their *accessibility*, through the transportation cost matrix; and (iii) the *heterogeneity* of producers through the regional supply schedule.[18]

The first two forces are gravitational in nature and were studied long ago by geographers and transport analysts who have proposed concepts, tools, and results known under the heading of “spatial interaction theory.” The aim is to study the formation of different types of flows, i.e., goods, people, and information, among places in response to localized supply and demand. To illustrate, consider a differentiated good traded across regions. Each region

[18] Cournot’s model, rediscovered and extended by Samuelson (1952), could have served as a basis for developing a trade theory with transportation costs, which came into being much later with the so-called new trade theories.

supplies a specific variety of the good, while individuals have a taste for variety and thus consume both the domestic and foreign varieties. If $y_{ij} \geq 0$ represents the export value from *i* to *j* and $y_{ii} \geq 0$ the consumption of the domestic varieties, region i's GDP Y_i is equal to the consumption of the domestic varieties plus the sum of its export values (or import values):

$$Y_i = \sum_j y_{ij} \qquad Y_j = \sum_i y_{ij} \ . \tag{9}$$

In the canonical model of spatial interaction theory developed by the English geographer Wilson (1970), the trade flow from region *i* to region *j* is given by the following *gravitational* law:

$$y_{ij} = K \frac{Y_i Y_j}{d_{ij}^{\theta}} , \tag{10}$$

which may be derived from the constrained maximization of a utility function that embodies a taste for variety, such as the CES, the translog, and the entropy, while being consistent with various trade settings (Head and Mayer, 2014). In (10), K is a parameter to be estimated, while d_{ij} is a reduced form that accounts for the impediments to trade from *i* to *j*. Note that d_{ij} need not be equal to d_{ji} while d_{ii} differs from zero because shipping variety *i* within region *i* is costly (e.g., Germany has a greater internal distance than Belgium). Unexpectedly, using the physical distance between regions *i* and *j* for d_{ij} works well in estimating (10). As a consequence, the function $d_{ij}^{-\theta}$ may be interpreted as a "spatial discount factor" that mimics the role played by transportation costs from *i* to *j* in Cournot's model and is thus a measure of the accessibility of *j* from *i*. A high value of the distance elasticity θ means that proximity is a crucial determinant of trade. At the other extreme, in a world where distance no longer matters, the parameter θ would be equal to zero.

Note that K must take on the following functional form:

$$K = \frac{Y}{\sum_i \sum_j Y_i Y_j d_{ij}^{-\theta}} ,$$

for (10) to be consistent with the constraints (9). As a consequence, (10) embodies multilateral, instead of bilateral, resistance terms. It was not until Anderson and van Wincoop (2003) that the importance of such general equilibrium effects was recognized by economists who accounted for them by adding regional fixed effects to the gravity equation.

Over the last decade, trade economists have successfully explored the microeconomic underpinnings of the gravity equation. They have also developed theory-consistent estimations of the spatial discount parameter by

taking into account a broad range of explanatory variables, such as prices, costs, and sophisticated measures of trade impediments, as well as heterogeneous firms and consumers. Through a meta-analysis of the various estimations of the distance elasticity θ in the literature, Head and Mayer (2014) find that the average distance elasticity is equal to 0.91. In other words, *on average, doubling the distance between two countries almost halves the volume of trade between these countries*. We are therefore far from a world in which distance would no longer be a dominant characteristic of the world economy. In fact, it is quite the opposite. Distance still dominates most aspects of international trade (Leamer, 2007; Brakman and van Marrewijk, 2008).

3.1.2 Transportation networks

By assigning different degrees of centrality to nodes in a transportation network, the specific network pattern favors some places at the expense of others. For example, transport infrastructure has been built in West and East Africa to allow former colonies to export their mineral resources to developed countries overseas. As this infrastructure is more efficient than the infrastructure that connects neighboring countries, it reduces the transport costs for imports from overseas more than for imports from neighboring trading partners. Bonfatti and Poelhekke (2015) 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 intranational transport networks designed during the colonial period still shape the intensity and nature of international trade flows. To a certain extent, the gravity equation would thus reflect the fact that connecting two neighboring countries is often cheaper and easier than two remote countries.

A fairly sizable literature in operations research and regional science studies how transportation activities affect commodity flows and the structure of the spatial economy (Thomas, 2002). For example, the development of new transportation methods has vastly changed the way in which distance affects transport costs over the last 200 years. This history is briefly as follows. The long period during which all movement was very costly and risky was followed by another during which, thanks to technological and organizational advances, ships could cross longer distances in one go, thus reducing their number of stops. On land, it was necessary to wait for the advent of the railroad for appreciable progress to occur, but the results were the same. In both cases, long-distance journeys became less expensive and no longer demanded the presence of relays or rest areas. Such an evolution in technologies has favored places of origin and destination at the expense of intermediate places. In other words, increasing returns in transport explain why places situated between large markets and transport nodes have lost many of their activities. As a consequence, *the construction of new and large transport infrastructures is beneficial to the main centers that the infrastructure connects, but not to the regions it goes through* (Beckman and Thisse, 1986). Having this in mind, it is hardly a shock that not much happened in those transit regions, despite the high expectations in the local populations.

The traditional problem of firm location theory is to seek—for some arbitrary but fixed quantities to be shipped to and from an arbitrary but given set of locations on the network—the place with the lowest total transportation costs. The above argument can be used to show that this problem may be reduced from an infinite number of alternatives to a finite number with one optimal location. This finite set is formed by the transport nodes and the input/output markets. Such a characterization has a great deal of intuitive appeal in spatial economics, as it shows that not all connected locations compete on an equal footing. What is more, it reveals the discontinuous nature of firm relocation. In other words, *substitution between locations does not occur in the small but in the large.* The shape of the transport network thus has an impact on firms' locational choices and trade flows. Of course, we face here the problem of simultaneity discussed in Section 2: the network affects locational choices and the nature of trade flows, but new links are built because the size of trade flows between two places requires extra capacity, which in turn reinforces the attractiveness of the two nodes.

3.2 Market access and firms' location

We now turn our attention to NEG in which the mobility of goods and of production factors is equally important, an approach praised by Ohlin. When compared with earlier attempts made in regional economics, an appealing feature of NEG is its very strong connections with several branches of modern economics.

3.2.1 The home-market effect

The neoclassical theory of the mobility of production factors and goods predicts a market outcome in which production factors receive the same reward regardless of where they operate. Indeed, when each region is endowed with the same production function that exhibits constant returns to scale as well as a decreasing marginal productivity, capital 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. Therefore, the free mobility of goods and capital would guarantee the equalization of wages and capital rents across regions and countries. In this case, the size of markets would be immaterial to people's welfare.

However, we are far from seeing such a featureless world. To solve this contradiction, NEG takes a radical departure from the standard setting. NEG assumes that the main reason why there is no convergence is that firms do not operate under constant returns but under *internal* increasing returns.[19] This point was made by Krugman

[19] The shift from external to internal returns may be explained by the difference in the spatial scale. Knowledge and informational spillovers (discussed in Section 2) are very localized, making them effective at the local level. But they probably cease to play a role at the interregional level where distances are much greater.

(1980) in a paper now famous because it highlights how *market size* and *market accessibility* interact to determine the location of an industry. The idea that size matters for the development of a region or country was emphasized by the economic historian Pollard (1981) for whom "it is obviously harder to build up an industrial complex without the solid foundation of a home market." In contrast, economic integration and regional trade agreements lower the importance of domestic markets and allow small regions and countries to supply larger markets.

Both economists and geographers agree that a large market tends to increase the profitability of the firms established there. In his famous location problem, where a firm chooses the location that minimizes its total transport costs, Weber (1909) showed that the market—or input source—with a weight exceeding the weighted sum of the other markets and input sources is always the firm's optimal location. More generally, the idea is that locations with good access to several markets offer firms a greater profit because these locations let firms save on transportation costs and lower their average production cost by selling a bigger output. In sum, firms would seek locations with the highest market potential where demand is high and transport costs are low. Most empirical works use the concept of *market potential*, introduced by the American geographer Harris (1954) and defined as the sum of regional GDPs weighted by the inverse of the distance to the region in question where the sum includes the region itself and its internal distance, as a reduced-form expression derived from general equilibrium trade theory. Econometric studies suggest that market potential is a powerful driver of increases in income per capita (Mayer, 2008). In other words, larger and/or more centrally located regions or countries are richer than regions or countries with small local markets and few neighbors or neighbors that are also small.

Nevertheless, as firms set up in the large regions, competition is also heightened, thereby holding back the tendency to agglomeration. Indeed, revisiting Hoteling's (1929) pioneering work, d'Aspremont el al. (1979) show that price competition is a strong dispersion force. This has a far-fetched implication: everything being equal, *competition fosters the dispersion of firms*. By relaxing competition, *product differentiation allows firms to seek the most accessible location* (Fujita and Thisse, 2013). Consequently, the interregional distribution of firms producing a tradable good is governed by two forces that pull in opposite directions: the agglomeration force is generated by firms' desire for market access, while the dispersion force is generated by firms' desire to avoid market crowding. Thus, the equilibrium distribution of firms across regions can be viewed as the balance between these two opposing forces.

The intensity of the agglomeration force decreases with transport costs, whereas the dispersion force gets stronger through tougher competition between regions. Although it is the balance of these forces that determines the shape of the spatial economy, there is no clear indication regarding the relative intensity of those forces as transport costs decrease. This is why the main questions that NEG addresses keep their relevance: When do we observe an agglomerated or a dispersed pattern of production at the interregional level? What is the impact that

decreasing transport and trade costs have on the intensity of the agglomeration and dispersion forces operating at that spatial scale?

Location and market size. The standard model involves two regions (North and South) and two production factors (capital and labor). The global economy is endowed with K units of capital and L units of labor. Each individual is endowed with one unit of labor and K / L units of capital. Capital is mobile between regions and capital owners seek the higher rate of return; the share $\lambda \geq 1/2$ of capital located in North is endogenous. Labor is spatially immobile but perfectly mobile between the sectors; the share of workers located in North is exogenous and equal to $\theta \geq 1/2$. Both regional labor markets are perfect. Capital and labor are used by firms that produce a CES-differentiated product under increasing returns and monopolistic competition (Dixit and Stiglitz, 1977). Let $f > 0$ be the fixed capital requirement and $c > 0$ the marginal labor requirement needed for a firm to enter the market and produce one variety of the differentiated good. Capital market clearing implies that the number of firms is exogenous and given by K / f. Finally, shipping the differentiated good between the two regions is costly.

The above-mentioned system of push and pull reaches equilibrium when the capital return is the same in both regions. In this event, *North hosts a more-than-proportionate share of firms*, a result that has been coined the "home-market effect" (HME).[20] Since North is larger in terms of population and purchasing power, it seems natural that North should attract more firms than South. What is less expected is that the initial size advantage is magnified, that is, the equilibrium value of λ exceeds θ. What the HME shows is that the market-access effect dominates the market crowding effect. Since $(\lambda - \theta) K > 0$ units of capital move from South to North, capital does not flow from the region where it is abundant to the region where it is scarce.

How does a lowering of interregional transport costs affect this result? At first glance, one could expect the market-access effect to be weaker when transport costs are lower. In fact, the opposite holds true: more firms choose to set up in North when it gets cheaper to trade goods between the two regions. This somewhat paradoxical result can be understood as follows. On the one hand, lower transport costs makes exports to the smaller market easier, which allows firms to exploit their scale economies more intensively by locating in North; on the other hand, lower transport costs also reduce the advantages associated with geographical isolation in South where there is less competition. These two effects push toward more agglomeration, implying that, as transport costs go down, 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 transport of goods cheaper in both directions, the construction of new infrastructure may induce firms to pull out of the smaller region*. In other words, connecting lagging regions to dynamic urban centers may weaken their industrial base. This result may

[20] See Baldwin et al. (2003), Fujita and Thisse (2013) and Zeng (2014) for a discussion of the HME in different set-ups.

come as a surprise to those who forget that highways run both ways. What is more, the intensity of competition in domestic markets matters for trade. Since large markets tend to be more competitive, penetrating such markets is more difficult than exporting toward small regions, making the former regions even more attractive than the latter.

Wages and market size. Although it is convenient to assume equal wages across regions because this allows the impact of falling transportation costs to be isolated, the assumption clashes with casual evidence. How wages vary with firms' location is best studied in a full-fledged general equilibrium model where wages are endogenous. As firms congregate in the larger region, competition in the local labor market intensifies, which should lead to a wage hike in North. Since consumers in North enjoy higher incomes, local demand for the good rises and this makes North more attractive to firms located in South. However, the wage hike associated with the establishment of more firms in North generates a new dispersion force, which lies at the heart of many debates regarding the de-industrialization of developed countries, i.e., their high labor costs. In such a context, firms are induced to relocate their activities to South when the lower wages there more than offset the lower demand Takahashi et al. (2013) have shown that the equilibrium wage in North is greater than the equilibrium wage in South. Furthermore, the HME still holds. In other words, though the wage paid in North exceeds those paid in South, market access remains critical when determining the location of firms.

Furthermore, if the size of the larger region grows through the migration of workers from South to North, the interregional wage gap widens. Therefore, *fostering the mobility of workers could well exacerbate regional disparities*. Nevertheless, Takahashi et al. showed that the magnification of the HME discussed above no longer holds: as transport costs steadily decrease, both the equilibrium wage and manufacturing share first rise and then fall because competition in the larger labor market gets very strong. Therefore, market integration and factor mobility favor the agglomeration of activities within a small number of large regions.

It is commonplace in macroeconomics and economic policy to think of unemployment as a national problem, the reason being that labor market institutions and demographic evolutions are often country-specific. Yet, empirical evidence reveals the existence of a strong correlation between high (low) unemployment rates and a low (high) GDP per capita across regions belonging to the same EU country. This should invite policy-makers to pay more attention to the regional aspects of unemployment. In particular, is higher interregional labor mobility the right solution to large regional employment disparities? Not necessarily. As migrants get absorbed by the labor market of the core region, agglomeration economies come into play, which reduces the number of job seekers. Such a scenario is more likely to arise when migrants are skilled. In contrast, the opposite evolution characterizes the lagging region, which loses its best workers. Epifani and Gancia (2005) illustrate this contrasted pattern by introducing job search frictions à la Pissarides in a standard NEG setup and conclude that "migration from the

periphery to the core may reduce unemployment at first, but amplify them in the long run." Such a result clashes with the widely-spread idea that geographical mobility is *the* solution to regional unemployment disparities. Even though it would be daring to draw policy recommendations from a single theory paper, it is clear that more research is needed to understand fully the impact of labor mobility on the functioning of local labor markets when market size and agglomeration economies are taken into account.

***Heterogeneous firms*.** So far, we have assumed that firms are homogeneous. However, the evidence is mounting that firms differ vastly in productivity (Bernard et al., 2007). This is reflected in the firms' ability to compete in the international marketplace. For example, Mayer and Ottaviano (2007) observe that the top 1 percent of European exporters account for more than 45 percent of aggregate exports, while the top 10 percent of exporting firms account for more than 80 percent of aggregate exports. In short, a few firms are responsible for the bulk of exports. Having such numbers in mind, it is thus legitimate to ask what the HME is when firms are heterogeneous and also when they are, or are not, sorted out across regions according to their productivity.

We have seen in Section 2 that heterogeneous workers are sorted between cities along educational lines. A comparable process is at work in the case of heterogeneous firms: *the more productive firms locate in the larger region*, whereas the less productive firms seek protection against competition by setting up in the smaller region (Nocke, 2006; Okubo et al., 2010). Furthermore, despite the greater competition in North, the HME still holds. Nevertheless, the mechanism that selects firms differs from the sorting of workers. Indeed, the gathering of the more productive firms renders competition very tough in North, which leads the inefficient firms to locate far apart to avoid the devastating effects of competition with efficient firms. This sparks a productivity gap between regions, which is exacerbated when the difference in size between regions increases. 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. This result is confirmed by the literature that follows Syverson. However, Combes et al. (2012) show that productivity differences explain only a small share of the urban wage premium. In other words, *agglomeration economies are stronger than selection effects*.

***Care is needed*.** First of all, the above results were obtained using specific models so the results' robustness remains an open question. Second, the share of the manufacturing sector has shrunk dramatically in developed economies. So one may wonder what the HME becomes when we consider the location of services that are often non-tradable. In this case, the HME still holds if North is sufficiently large to overcome the competition effect. Otherwise, the larger region no longer provides a sufficiently big outlet to host a more-than-proportionate share of firms. In this case, the smaller region accommodates a larger share of firms (Behrens, 2005).

Third, and last, the HME is studied in a two-region setting. Unfortunately, it cannot readily be extended to multi-regional set-ups because there is no obvious benchmark against which to measure the "more-than-proportionate"

share of firms.[21] A multi-regional setting brings about a new fundamental ingredient—the variability in regions' accessibility to spatially dispersed markets. In other words, the relative position of a region within the network of exchanges (which also involves cultural, linguistic, and political proximity) matters. Any global (local) change in this network such as market integration or the construction of major transportation links is likely to trigger complex effects that vary in non-trivial ways with the properties of the graph representing the transportation network (Behrens and Thisse, 2007). For example, in a multi-regional setting, the greater specialization of a few regions in one sector does not necessarily mean that this sector becomes more agglomerated, and vice versa. Therefore, it is hardly a shock 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, 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 are selected randomly (e.g., hub and spoke networks). For this, one needs a mathematical framework that is tractable but yet rich enough to analyze meaningful effects. Working with a NEG model that encompasses asymmetric regions, costly trade, and transport tree-networks generated randomly, Barbero et al. (2015) confirm that local market size (measured by population) and accessibility (measured by centrality in the trading network) are crucial in explaining a region's wage; they also confirm that local market size (measured by industry expenditure share) explains well the location of firms. Using Spanish data and computed transportation costs, Barbero et al. find that the model is good at predicting the location of industries but less accurate concerning the spatial pattern of wages. The authors also observe that, after three decades of major road investments, the distribution of industries had not changed much in Spain. This might suggest that, once a few key connections exist, the supply of transportation links obeys the law of decreasing return.

The second method is to study empirically the causality between market access and the spatial distribution of firms. There is a wealth of evidence suggesting that market access is associated with firms' location, higher wages, and employment. 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." For example, one of the

[21] So far, the best that has been accomplished in a multi-regional setting is Behrens et al. (2009) but they assume that wages are equal across regions. Zeng and Uchikawa (2014) show that the wage ranking is the same as the market size ranking. However, they assume that transport costs are the same between any pair of regions.

more remarkable geographical concentrations of activities is what is known as the "manufacturing belt" that accommodated around four-fifths of the U.S. manufacturing output for a century or so within an area that was one-sixth of the country's area. Klein and Crafts (2012) conclude that "market potential had a substantial impact on the location of manufacturing in the USA throughout the period 1880–1920 and ... was more important than factor endowments." In the same vein, Head and Mayer (2011) summarize their analysis of the relationship between market proximity and economic development over 1965–2003 by saying that "market potential is a powerful driver of increases in income per capita."

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 transportation costs. In other words, *even at the interregional level, proximity matters*, but the reasons for this are not the same as those uncovered in Section 2. However, both sets of results hinge on the same principle: *small initial advantages may be translated into large ex post advantages once firms operate under* (external or internal) *increasing returns*.

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, the two regions are supposed to be the same size and the internal fabric of each region (e.g., the magnitude of agglomeration economies) determines the circumstances in which a region accommodates the larger number of firms. Second, workers are allowed to migrate from one region to the other, thus leading to some regions being larger than others. The former case—when the two regions are a priori identical—is studied below, while the latter case is investigated in Section 3.3 because the mobility of labor generates effects that differ from those observed under the mobility of capital.

3.2.2 Why do asymmetric industrial clusters emerge in a symmetric world?

According to Porter (1998), the formation of industrial clusters depends on the relative strength of three distinct forces: the size of intrasectoral agglomeration economies, the intensity of competition, and the level of transport costs. Despite the existence of a huge empirical—and inconclusive—literature devoted to industrial clusters, how the three forces interact to shape the regional economy has been neglected in NEG, probably because working with a model that accounts for the main ingredients of urban economics and NEG seems out of reach. Yet the formation of clusters can be studied by adopting a "reduced-form" approach in which a firm's marginal production cost in a region decreases with the number of firms locating in the region. In doing this, one captures the effect of agglomeration economies, such as those discussed in Section 2, and can study how agglomeration economies operating at the local level interact with the dispersion force generated by market competition in the global economy through lower trade costs (Belleflamme et al., 2000). In a spatial equilibrium, firms earn the same profits.

However, if firms observe that one region offers higher potential profits than the other, they want to move to that region. In other words, the driving force that sustains the relocation of firms is the profit differential between North and South.

To show why and how a hierarchy of clusters emerges, we look at the interplay among the above three forces as a *symmetry-breaking device.* Therefore, we start with a perfectly symmetric set-up in which firms and consumers are evenly dispersed between North and South. When trade costs start decreasing, trade flows grow but, in the absence of agglomeration economies, firms stay put because spatial separation relaxes the competition between firms. Things are very different when agglomeration economies are at work. In this case, when trade costs fall enough, some firms choose to produce in North, say, instead of South in order to benefit from a lower marginal cost while maintaining a high volume of export. As trade costs keep decreasing, a growing number of firms choose to set up in North where the marginal cost decreases further. Note that firms tend to gather in one region despite the fact that the two markets where they sell their output have the same size. What now drives firms' agglomeration is no longer market size but the endogenous level of agglomeration economies.

But where does agglomeration occur? Will it be in North or in South? Consider an asymmetric shock that gives a region a small initial advantage. If this shock remains fixed over a long period, firms will attune their behavior accordingly. The region benefiting from the shock, however small, will accommodate the larger cluster. Hence, *regions that were once very similar may end up having very different production structures* as market integration gets deeper. Once more, lowering trade costs drives the economy toward more agglomeration in one region at the expense of the other.

Are growing regional disparities necessarily bad in this context? The answer is no. A planner whose aim is to maximize global efficiency sets up more asymmetric clusters than what the market delivers. To explain—at the first-best optimum, prices are set at the marginal cost level while locations are chosen to maximize the difference between agglomeration economies and transport costs. In contrast, at market equilibrium, firms take advantage of their spatial separation to relax price competition and do not consider the positive externalities associated with their location decision. So the optimal configuration tends to involve a more unbalanced distribution of firms than the market outcome. If agglomeration economies become increasingly important in some sectors, their uneven geographical distribution need not signify a wasteful allocation of resources. On the contrary, *the size of the clusters could well be too small*. However, the region with the larger cluster benefits from lower prices through larger agglomeration economies, more jobs, and a bigger fiscal basis.

3.3 Labor mobility

The mobility of capital and the mobility of labor do not obey the same rules. First, while the movement of capital to a region brings with it production capability, the returns to capital need not be spent in the same region. In

contrast, when workers move to a new region, they bring with them *both their production and consumption capabilities* (putting aside remittances). As a result, migration affects the size of the labor *and* the product markets in both the origin and destination regions. Second, while the mobility of capital is driven by differences in *nominal* returns, workers care about their *real* wages. In other words, differences in living costs matter to workers but not to capital owners.

3.3.1 The core-periphery structure

The difference in the consequences of capital and labor mobility is the starting point of Krugman's celebrated 1991 paper that dwells on the idea that the interregional economy is replete with *pecuniary externalities* generated by the mobility of workers.[22] Indeed, when some workers choose to migrate, their move affects the welfare of those who stay behind because migration affects the size of the regional product and labor markets. 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.

To achieve his goal, Krugman (1991) considers the classical 2 x 2 X 2 setting of trade theory. There are two goods, two types of labor, and two regions. The first type of labor (workers) is mobile and the only input in the first (manufacturing) sector, which operates under increasing returns and monopolistic competition; shipping the manufactured good is costly. The second type of labor (farmers) is immobile and the only input in the second (farming) sector, which produces a homogeneous good under constant returns and perfect competition; shipping the agricultural good incurs no cost. What drives the agglomeration of the manufacturing sector is the mobility of workers. For this, Krugman considers a setting in which both farmers and workers are symmetrically distributed between North and South and asks when this pattern ceases to be a stable spatial equilibrium.

Two main effects are at work: one involves firms and the other, workers. Assume that North becomes slightly bigger than South. At first, this increase in market size leads to a higher demand for the manufactured good, thus attracting more firms. The HME implies that the hike in the number of firms is more than proportional to the increase in market size, thus pushing nominal wages upward. In addition, the presence of more firms means that

[22] It is worth noting that Krugman's paper and that of Glaeser et al. (1992), which spark the resurgence of urban studies in mainstream economics, were published during the same period in the *Journal of Political Economy*. Note also that Romer (1990) precedes Krugman by one year. Both authors use the CES model of monopolistic competition. This concomitance and convergence of ideas is hardly a shock. Indeed, explaining technological progress and urban agglomerations can hardly be handled using the perfect competition – constant returns paradigm.

a greater number of varieties are produced locally and therefore prices are lower in North. As a consequence, real wages rise so that North should attract a new flow of workers. Therefore, there is *circular causation* à la Myrdal in which these two effects reinforce each other. This snowball effect seems to lead inevitably to the agglomeration of the manufacturing sector in North that becomes the *core* of the global economy.

But the snowball may not form. Indeed, the foregoing argument ignores several other effects triggered by the migration of workers. On the one hand, the increased supply of labor in North tends to push wages downward. On the other hand, since new workers are also consumers, there will be a hike in local demand for the manufactured good, which leads to a higher demand for labor. This is not yet the end of the story. As more firms enter the local market, there is increased competition to attract workers so the final impact of migration on nominal wages is hard to predict. Likewise, there is increased competition in the product market as well as greater demand. Combining these various effects might well lead to a "snowball meltdown," which results in the spatial dispersion of firms and workers.

Krugman's great accomplishment has been to integrate all these effects within a single framework and to determine precisely the conditions under which the above prediction holds or not. Starting from an arbitrarily small difference between regions, Krugman singled out the cases in which there is agglomeration or dispersion of the manufacturing sector. He showed that the value of transport costs is again the key determining factor. If transport costs are sufficiently high, the interregional shipment of goods is low. In this event, firms focus on regional markets. Thus the global economy displays a symmetric regional pattern of production. In contrast, when transport costs are sufficiently low, then all manufacturers will concentrate in North; South will supply only the agricultural good and will become the *periphery*. In this way, firms are able to exploit increasing returns by selling more in the larger market without losing much business in the smaller market. Again, lowering trade costs fosters the gathering of activities. The core-periphery model therefore allows for the possibility of *convergence* or *divergence* between regions, whereas the neoclassical model based on constant returns and perfect competition in the two sectors predicts only convergence. Consequently, Krugman presented a synthesis of the polarization and neoclassical theories. His work appealed because the regional disparities associated with the core-periphery structure emerge as a *stable equilibrium* that is the involuntary consequence of decisions made by a large number of economic agents pursuing their own interests.

When agents are mobile, supply and demand schedules are shifted up and down by the agents' relocation across places. It is no surprise, therefore, that it is not possible to come up with an analytical solution of the core-periphery model. This is what led Krugman to resort to numerical analysis to uncover the impact of decreasing transport costs on the location of economic activity. 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.

Krugman's paper triggered a huge flow of research. The best synthesis of what has been accomplished in NEG remains Baldwin et al. (2003).

Despite its great originality, the core-periphery model has several shortcomings. The following list, while not exhaustive, covers a fair number of issues. (i) The model overlooks the various congestion costs and agglomeration economies generated by the concentration of activities, discussed in Section 2. (ii) It only accounts for two sectors and two regions. (iii) The agricultural sector is given a very restricted role, its job being to guarantee the equilibrium of the trade balance. Along the same line, it is hard to see why trading the agricultural good costs nothing in a model seeking to determine the overall impact of trade costs. All these features have attracted a lot of attention, but the "dimensionality problem" is the most challenging one.

Having said that, we must stress the work by Helpman (1998) who argued that decreasing freight costs may trigger the dispersion, rather than the agglomeration, of economic activities when the dispersion force lies in the supply of non-tradable services (housing) rather than immobile farmers. In this case, the various congestion effects discussed in Section 2 put a brake on the agglomeration process, and thus *Krugman's prediction is reversed*.[23] The difference in results is easy to understand. Commuting and housing costs rise when consumers join the larger region/city, which strengthens the dispersion force. Simultaneously, lowering transport costs facilitates interregional trade. By combining these two forces, we see why dispersion arises. By neglecting that agglomeration of activities typically materializes in the form of cities where competition for land acts as a strong dispersion force, the core-periphery model remains in the tradition of trade theory. Therefore, conclusions drawn from this model are, at best, applicable to very large areas.

3.3.2 Is technological progress an agglomeration force?

Krugman's core-periphery model highlights the role of market integration as the main force driving the location of economic activity. This agrees with classical location theory in which firms aim to minimize transportation costs. Though relevant, market integration is unlikely to be the sole force shaping the economic landscape at the interregional level. This state of affairs has led Tabuchi et al. (2015) to revisit NEG by focusing on technological progress in the manufacturing sector. In addition, these authors recognize that workers are imperfectly mobile. Indeed, migration generates substantial non-pecuniary costs created by differences in languages/dialects, cultures, and religions within and between nations, which have a lasting influence on individual well-being. Temporary and return migration is evidence that migrants bear permanent social dislocation costs when they live away from their place of origin. Such costs explain the low mobility of European workers.

[23] This idea was already known to Weber (1909): "These deglomerative factors all follow from the *rise of land values*, which is caused by the increase in the demand for land, which is an accompaniment of all agglomeration.

Findings differ in various respects from those obtained in the core-periphery model. First, in Krugman (1991) and followers, the incentive to move shrinks as trade costs fall because prices and nominal wages converge. What drives Krugman's result is the change in the sign of the utility differential when trade costs fall below some threshold. Note, however, that the absolute value of the utility differential steadily decreases with trade costs. In contrast, even when North is slightly bigger than South, an exogenous technological progress that reduces the labor marginal requirement in the two regions makes North more attractive by increasing the wages and decreasing the prices that prevail in this region. In other words, technological progress tends to exacerbate differences between the two regions and thus raises the incentive to move from South to North. Another major difference is worth pointing out. Falling trade costs fosters 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.

Second, workers move to North 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. Thus, high mobility costs lower the productive efficiency of the European economy but avoid increasing regional disparities.

Third, once it is recognized that workers are heterogeneous, those who move from South to North are the most productive ones. Indeed, those workers will benefit most from the hike in the price of one efficiency unit of labor, and they also have the lowest mobility costs. This affects the two regions in opposite ways: North becomes more productive, whereas South loses its best workers. Phrased differently, there is spatial sorting of workers across regions, very much as there is sorting of workers across cities. As technological progress steadily develops, the stock of human capital in North rises faster than in South, where it may even decline. As a consequence, regional disparities get deeper, regardless of the level of transportation costs. This does not mean that South is necessarily trapped in stagnation or decline.

First, as expected, differences in human capital endowments affect the economic development of the EU regions. However, Rodriguez-Pose and Vilalta-Bufia (2005) accurately stress that human capital may not deliver its full impact if human resources are left idle or poorly used. Factors like the matching between the regional educational supply and labor demand, as well as the satisfaction of employers with the skills of their workers and of employees with their capacity to use their training, though often neglected in measures of human capital, play a significant role at the local level. Second, as seen in Section 2, not all sectors and activities benefit from agglomeration economies. Therefore, by offering cheaper land and the required skills, lagging regions may attract such activities. Note that labor productivity is often lower in poorer regions than in the richer core. If wages are influenced by

factors at the national level, such as wage bargaining between trade unions and employers in several member countries of the EU, workers in the lagging regions may be priced out of the market (Faini, 1999).

3.4 Does the market yield over- or under-agglomeration?

Whether there is too much or too little agglomeration is unclear. Yet, speculation on this issue has never been in short supply and it is fair to say that this is one of the main questions that policy makers would like to address. Contrary to general beliefs, the market need not lead to the over-agglomeration of activities as competition is a strong dispersion force. We have discussed above two basic mechanisms that may outweigh this force and lead to the spatial clustering of activities. The former is the HME, which points to the relative agglomeration of firms in the large regions. The latter is related to the joint concentration of firms and workers in a few regions to form big markets. Since the mobility of capital and labor is driven by different forces, there is no reason to expect the answer to the question that serves as the title of this subsection to be the same.

3.4.1 Does the home market effect generate excessive agglomeration?

Firms want to reconstitute their markups by locating in spatially separated markets. However, other forces may outweigh this effect and this leads to the concentration of firms in a few regions. When firms move from one region to another, they impose negative pecuniary externalities on the whole economy. More precisely, firms neglect the impact of their move on product and input markets in both destination and origin regions. The social surplus is lowered because location decisions are based on relative prices that do not reflect the true social costs. However, the inefficiency of the market outcome does not tell us anything about the excessive or insufficient concentration of firms in the big regions. In fact, *the HME involves too many firms located in the larger region*. The intuition is easy to grasp. A profit-maximizing firm chooses the location that minimizes the transportation costs it bears to serve foreign markets. Therefore, since firms absorb more freight when exporting from the smaller to the larger region than vice versa, firms are incentivized to locate in the larger region. Tougher competition there holds back the agglomeration process, but this dispersion force is not strong enough for a sufficiently large number of firms to set up in the smaller region. However, it is worth noting that the first-best distribution of firms still involves a share of firms exceeding the size of the larger region (Ottaviano and Van Ypersele, 2005).

3.4.2 Is the core-periphery structure inefficient?

Thus far, NEG has not been able to provide a clear-cut answer to this fundamental question. However, a few results seem to show some robustness. In the core-periphery model, the market outcome is socially desirable when transport costs are either high or low. While in the former case activities are dispersed, in the latter they are agglomerated. In contrast, for intermediate values of these costs, the market leads to the over-agglomeration of the manufacturing sector (Ottaviano and Thisse, 2002). Furthermore, when transport costs are sufficiently low, agglomeration is preferred to dispersion in the following sense: people in the core regions can compensate those

staying in the periphery, whereas those staying in the periphery are unable to compensate those workers who choose to move to what becomes the core regions (Charlot et al., 2006). This suggests that interregional transfers could be the solution for correcting regional income disparities. It is worth stressing that such transfers do not rest here on equity considerations, but only on efficiency grounds. However, implementing such transfers, paid for by those who reside in the core regions, may be politically difficult to maintain in the long run. In addition, they may give rise to opportunistic behavior in the periphery.[24]

Tackling this issue from a dynamic perspective sheds additional light on the problem. It has been argued for long that growth is localized, the reason being that technological and social innovations tend to be clustered while their diffusion across places would be slow. For example, Hirschman (1958) claimed that "we may take it for granted that economic progress does not appear everywhere at the same time and that once it has appeared powerful forces make for a spatial concentration of economic growth around the initial starting points" while Hohenberg and Lees (1985) similarly argued that "despite the rapid growth of urban industries in England, Belgium, France, Germany and northern Italy after 1840 or so, economic development was a spatially selective process. Some regions deindustrialized while others were transformed by new technologies."

Fujita and Thisse (2003, 2013) revisit the core-periphery model in a setup combining NEG and endogenous growth theory and show that the growth rate of the global economy positively depends on the spatial concentration of the R&D sector. Furthermore, the core-periphery structure in which both the R&D and the manufacturing sectors are agglomerated is stable when transportation costs are sufficiently low. Such a result gives credence to the idea that growth and agglomeration go hand in hand. The welfare analysis undertaken by these authors also supports the idea that the additional growth spurred by agglomeration may lead to a Pareto-dominant move: when the growth effect triggered by the agglomeration of the R&D sector is strong enough, even those who live in the periphery are better off than under dispersion.

It is worth stressing that this Pareto-optimal move does not require any interregional transfer: it is a pure outcome of market interaction. However, the gap between the unskilled who live, respectively, in the core and in the periphery enlarges. Put differently, the rich get richer and so may do the poor, but without ever catching up. The welfare gap between the core and the periphery arises because of the additional gains generated by a faster growth that the skilled are able to spur by being agglomerated. This in turn makes the unskilled residing in this region better off, even though their productivity is the same as the one of those living in the periphery.

[24] Things become more complex when crowding effects arise when activities get agglomerated in a region. In this case, depending on the parameter values of the economy the equilibrium may yield either suboptimal agglomeration or suboptimal dispersion (Ottaviano et al., 2002; Pflüger and Südekum, 2008). Recall that crowding is more likely to arise in small than in large regions.

3.5 Vertical linkages and the spatial fragmentation of the supply chain

The econometric analysis undertaken by Crozet (2004), together with the observations made in Section 1, suggests that the low mobility of European workers makes the emergence of a Krugman-like core-periphery structure within the EU very unlikely. Therefore, 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 such as the EU. A second shortcoming of the core-periphery 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 intermediate goods.[25]

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

So far, agglomeration has been considered the outcome of a circular causation process fed by the mobility of workers. However, agglomeration of economic activities also arises in contexts in which labor mobility is very low, as in most European countries. This underscores the need for alternative explanations of industrial agglomeration. One strong contender is *the presence of input-output linkages between firms*: the output of one firm can be an input for another, and vice versa. In such a case, the entry of a new firm in a region not only increases the intensity of competition between similar firms; it also increases the market of upstream firm-suppliers and decreases the costs of downstream firm-customers. 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*. Indeed, when firms belonging to the final sector are concentrated in a single region, the local demand for intermediate inputs is very high, making this region very attractive to firms producing these intermediate goods. Conversely, because intermediate goods are made available at lower prices in the core region, firms producing final goods find that region very attractive. Thus, a cumulative process may develop that leads to industrial agglomeration within the core region.[26]

In this alternative setting, new forces arise. Indeed, if firms agglomerate in a region where the supply of labor is inelastic, then wages must surely rise. This in turn has two opposite effects. On the one hand, *consumers' demand*

[25] Intermediate goods represent 56 percent of total trade in goods, while final consumption goods only represent 21 percent of total trade in goods (OECD, 2009).

[26] Toulemonde (2006) has identified another mechanism of agglomeration, which bears some strong resemblance to Krugman and Venables (1995). When workers are a priori unskilled and immobile, some of them may choose to become skilled in order to be able to work in the manufacturing sector. As a result, they earn a higher income and, therefore, have a higher demand for manufacturing goods, making their region a larger and more attractive market to firms. At the same time, the installation of new firms within this region gives a stronger incentive to workers to improve their skill. As above, we obtain a mechanism of cumulative causation in which spatial mobility is replaced by sector-based mobility. In this case, income differences reflect the uneven distribution of human capital across regions. This approach to regional inequality is in the spirit of what we have discussed in Section 2.

for the final product increases because they have a higher income. This is again a market expansion force, now triggered by higher incomes rather than larger populations. On the other hand, such wage increases also generate a dispersion force. When the wage gap between the core and the periphery becomes sufficiently large, some firms will find it profitable to relocate to the periphery, even though the local demand for their output is lower than in the core. This is especially true when transport costs are low, because asymmetries in demand will then have a weaker impact on profits.

The set of equilibrium patterns obtained in the present setting is much richer than in the core-periphery model. In particular, if a deepening of economic integration triggers the concentration of industrial activities in one region, then, beyond a certain threshold, an even deeper integration may lead to a reversal of this tendency. Some firms now relocate from the core to the periphery. In other words, the periphery experiences a process of *re-industrialization* and, simultaneously, the core might start losing firms, thus becoming *de-industrialized*. As Fujita et al. (1999) put it, "declining trade costs first produce, then dissolve, the global inequality of nations."

Therefore, 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 (Williamson, 1965; Puga, 1999). The existence of such a curve may be obtained in several extensions of the core-periphery model—surveyed in Fujita and Thisse (2013)—and seem to be confirmed by several empirical and historical studies.[27] However, owing to differences in data, time periods and measurement techniques, it is fair to say that the empirical evidence is still mixed (Combes and Overman, 2004). Furthermore, this self-correcting effect can take too long in the face of some regions' urgent economic and social problems and the time horizon of policy-makers, which leads them to look for policies whose effects are felt more rapidly.

Note that the following coordination failure may prevent the redistribution of activities: many prices are not known in advance in South. Lack of adequate information may then prevent the development of a network of service and intermediate goods suppliers which leads to a vicious circle and persistent underdevelopment. In the presence of external effects, this problem is particularly acute. One solution is to have an agent who "internalizes" the various costs and benefits arising during the first stages of the take-off process who plays an entrepreneurial role that facilitates individual decisions, so that a cluster in South can form en masse.

3.5.2 Communication costs and the relocation of plants

A major facet of the process of globalization is the *spatial fragmentation* of the firm associated with vertical investments. Vertical investments arise when firms choose to break down their production process into various stages spread across different countries or regions. Specifically, the modern firm organizes and performs discrete

[27] See Barrios and Strobl (2009), Combes et al. (2011) and references therein.

activities in distinct locations, which together form a *supply chain* starting at the conception of the product and ending at its delivery. This spatial fragmentation of the firm aims to take advantage of differences in technologies, factor endowments, or factor prices across places (Feenstra, 1998). We now turn our attention to this problem.

Besides transportation costs, spatial separation generates another type of spatial friction, namely "communication costs." Indeed, coordinating activities within the firm is more costly when the headquarters and its production plants are physically separated because the transmission of information remains incomplete and imperfect. Furthermore, more uncertainty about production plants' local environment is associated with conducting a business at a distance. Again, this implies higher coordination costs, hence higher communication costs between the headquarters and its plants. In the same vein, monitoring the effort of a plant manager is easier when the plant is located near the headquarters than across borders. Lower communication costs make the coordination between headquarters and plants easier and therefore facilitate the process of spatial fragmentation.

For the international/interregional fragmentation of firms to arise, the intra-firm coordination costs must be sufficiently low so the operation of a plant at a distance is not too costly, whereas transportation costs must decrease substantially to permit the supply of large markets at low delivery costs from distant locations. To make low-wage areas more accessible and attractive for the establishment of their production, firms need the development of new information and communication technologies, as well as a substantial fall in trade costs. In this case, a certain number of firms choose to go *multinational*, which means that their headquarters are located in prosperous areas where they find the skilled workers they need while their plants are set up in low-wage areas, whereas the other firms remain spatially integrated (Fujita and Thisse, 2006; Spulber, 2007).

Manufacturing firms long ago started to relocate their production plants to regions where labor and land are cheaper than in large cities (Henderson, 1997; Glaeser and Kohlhase, 2004). However, transportation and communication costs for a long time imposed a limit to the distance at which plants could operate. The ongoing revolution in information and communication technologies freed some firms from this constraint, thus allowing them to move their plants much further away to countries where wages are much lower than in the peripheral regions where they used to establish their plants.[28] Hence, the following question: *Which "South" can accommodate firms' activities that are being decentralized?*

3.6 Do EU interregional transport policies fulfil their role?

In a way, this question may seem odd as the absence of good transport infrastructure is known to be one of the main impediments to trade. This is why international organizations such as the European Commission and the

[28] In recent years, there has been a modest reversal of outsourcing toward insourcing.

World Bank have financed a large number of transportation projects. As the key objective of the EU is deeper market integration among member countries, the construction of efficient and big transport infrastructures was seen as a necessary step toward this goal. However, this does not mean that one should keep increasing the supply of transport infrastructure.

In the EU, transport policy serves two main objectives. The first is to decrease trade costs as the aim is to build the EU internal market. The second objective is to promote the economic development and structural adjustment of lagging region. Arbitrage possibilities arising from competition and factor mobility are expected to induce a more than average growth performance in lagging regions. Having the economic engine in a higher gear would eventually make these regions reach the standard of living realized elsewhere. Where convergence does not set in swiftly, an insufficient stock of public infrastructure is often blamed. The EU and national governments have responded by pouring huge quantities of concrete in lagging regions.

The policy intervention also involved the design of pricing and regulation policies for interregional transport. All this has led to a strong increase in the volume of both freight and passenger transport. Nevertheless, national transport policies still depend on member countries. Using a NEG set-up in which transport costs between regions of the same country differ from trade costs between countries, Behrens et al. (2007) show that the welfare of a country increases when its internal transport costs are lowered because domestic firms increase their market share at the expense of foreign firms, while the trading partner is affected adversely for the same reason. As a consequence, we have something like a “fortress effect” in that accessing the increasingly integrated national market becomes more difficult, which may generate conflicts of interest between member countries.

The EU has sent rather mixed signals in terms of transport policy. In the first phase, the integration of markets for goods was the priority; later, the emphasis shifted to environmental and resource efficiency. As a result, the development of rail and waterways was favored over road and air transport. Yet road freight transport in the EU remains by far the dominant mode; the EU has a very different modal split from that in the U.S. International freight in the EU relies on road transport for 45 percent of traffic, on sea transport for 37 percent, on rail transport for 11 percent, and on inland waterways and pipeline transport for the remainder. In the U.S., rail transport (41 percent) is more important than road transport (32 percent), followed by pipeline (15 percent) and inland waterways. International passenger transport inside the EU also has a different modal split from that in the U.S. The U.S. relies on car and air transport, while the EU also relies on high-speed rail (HSR) transport. Thus, in the U.S., rail has an important share of the freight market while, in Europe, rail is more important for the passenger market.

3.6.1 The economic impacts of transport infrastructures

The economic performance of transport infrastructure can be improved by investing in transport infrastructure, by selecting investments more carefully, and by using the existing infrastructure better. Whether interregional transport infrastructure is beneficial in terms of welfare and whether it generates economic growth at the macroeconomic level are two different issues.

Assessing the benefits of transport investments ex ante, but also ex post, is difficult. There are two reasons for this. First, transport investments have a multitude of effects. They reduce trade barriers and so affect the pattern of trade, for freight as well as for services (via lower costs for business and tourism trips). As seen above, the outcome of a transport investment is a priori difficult to predict in a world where economic activities are increasingly footloose (see 3.4.1). The second difficulty is that the effect of an investment is ex post difficult to evaluate. The main reason is that there is no obvious counterfactual. A transport investment is often located where decision makers expect it to give rise to the largest benefits. But then, it becomes unclear whether it is the transport investment itself or the favorable pre-conditions that are at the origin of the observed effects.

The performance of transport infrastructure being an empirical question, we have chosen to discuss both ex ante and ex post methods. In particular, as in the urban economics section, we consider three approaches: the econometric approach, the model-simulation approach, and the case-study approach.

***Assessing transport investments by means of econometric models*.** In the post-Reagan period, public investments were expected to stimulate economic growth. In an influential paper, Aschauer (1989) used a reduced-form estimation and found high rates of return for public investments. This was the start of a series of macroeconomic studies that produced fairly mixed evidence about the impact of transport investments on national growth (Gramlich, 1994). Melo et al. (2013) conducted a meta-analysis of the existing empirical literature on the output elasticity of transport infrastructure. They showed that the productivity effects of transport infrastructure vary substantially across industries, tend to be higher in the US than in the EU, and are higher for roads compared to other transport modes of transport. The variation in estimates of the output elasticity of transport is also explained by differences in the methods and data used in the various studies. Failing to control for unobserved heterogeneity and spurious associations tends to result in higher values, while failing to control for urbanization and congestion levels leads to omitted variable bias. In addition, Puga (2002) highlighted several pitfalls of an aggregate approach. First, it could well be that transport investments happen just because economic growth allows the government to spend more money on infrastructure, not the other way around. Second, the first links of a transportation network could well be very productive, whereas the productivity of additional links decreases strongly.[29]

[29] A study by Combes and Lafourcade (2005) sheds light on that point. Based on the French road and highway network, these authors propose a measure of transportation costs that encompasses the characteristics of the network, vehicle and energy used,

Redding and Turner (2015) develop a general equilibrium framework in the spirit of Helpman-Tabuchi to assess the effects of transport investments on the location of production and population, as well as on variables like wages and prices. This allows these authors to construct the necessary counterfactuals needed to assess the effects of new transport investments. They find only limited evidence on the effect of interregional investments in the EU. Ahlfieldt and Feddersen (2015) study the impact of HSR on a corridor in Germany by comparing the effects on smaller towns with a HSR stop and those without a HSR stop. They find that as HSR decrease the cost of human interaction but trade costs remain unchanged, this type of projects has another effect on the core-periphery balance. Peripheral regions tend to experience negative effects through projects that reduce freight costs via a trade channel, as in NEG, but could benefit via Marshallian externalities from HSR projects.

Comparing the impact of transport investments in different non-EU parts of the world, Redding and Turner find that, across a range of countries and levels of development, new transportation infrastructures seem to generate similar effects. First, population density falls between 6 and 15 percent with a doubling of the distance to a highway or railroad, while highways decentralize urban populations and, to a less extent, manufacturing activity. Second, different sectors respond differently to different transportation modes. Another forceful piece of evidence is Faber (2014) who showed that the construction of new highways in China decreased trade costs, but, as suggested by NEG, re-enforced the core cities at the expense of the periphery.

One limitation of the econometric assessment approach is that transport investments are chosen by a political process, which can lead to the selection of poor investments. For example, Knight (2002) has found that, for the U.S. Federal Highway Fund, about half of the investment money was wasted. Therefore, any econometric ex post assessment has the tough task of distinguishing between poor political selection mechanisms and the potential effects of a well-selected transport investment.

Assessing transport investments by means of simulations. When a reliable multi-regional simulation model is available, one can simulate the effects of transport investments and discriminate between the effects of the selection process and the productivity of a transport infrastructure. Only a handful of such models exist in the world.[30] To this end, the European Commission has developed a spatial computable general equilibrium model (SCGE), RHOMOLO, where different policy shocks can be simulated at the regional level to obtain an ex-ante

labor cost, taxes, and general charges paid by carriers. Combes and Lafourcade find that the 38 percent average decline in freight costs observed in France between 1978 and 1998 is mostly explained by technological improvements in trucking and the deregulation of the road transport industry. In contrast, the impact of infrastructure and fuel costs is weak. Thus, the efficiency of the interregional transport infrastructures that were last built seems, at best, marginal.

[30] See Bröcker and Mercenier (2011) for a survey of the literature.

impact assessment.[31] The spatial implications of the general equilibrium approach followed in RHOMOLO have been investigated by Di Comite and Kancs (2014), who describe how the main agglomeration and dispersion forces of NEG enter the model: agglomeration is driven by increasing returns to scale, the use of intermediate inputs and localized externalities, while dispersion is driven by costly trade and locally-produced varieties entering consumer utility asymmetrically (calibrated on observed trade flows). Capital and labor are mobile; and vertical linkages are accounted for using regionalized international input-output matrices. The model is implemented for the 267 NUTS-2 regions of the EU and used to assess the effect of investments that reduce trade costs and its properties are tested by simulating the impact of planned Cohesion Policy investments in infrastructure, whose main target are the poorer, peripheral regions. The aim of the exercise is to isolate the effect of the different economic mechanisms identified in subsections 3.2 and 3.3, for which three scenarios are simulated.

*Scenario 1: Isolating the effect of capital mobility***.** By switching capital mobility on and off, allowing savings in one region to be invested in other regions, the authors find that the tendency toward the equalization of the rates of return on investments spreads the growth effects of the transport investments more equally. This is the home-market effect at work: although the poorer (peripheral) regions received a larger share of the transport investments, the relocation of capital leads to more growth in the other EU regions.

Scenario 2: Isolating the effect of labor mobility. By switching labor mobility on and off, allowing workers to relocate where their real wages are higher according to estimated elasticities, the authors find that the region receiving the initial investment will benefit from a lower cost of living. This attracts more workers and increases the size of the region, its production and consumption, which should foster agglomeration. However, since consumer tastes are calibrated in each region based on the observed trade flows in the base year, the growing regions also demand more from the peripheral regions, which bids up prices and prevents a strong agglomeration effect. The cost-of-living effect is found to be stronger than the labor market crowding effect, thus magnifying the beneficial effect of local investments and making the lagging region better off, but the effect is much localized**.**

Scenario 3: Isolating the effect of vertical linkages. By switching inter-regional consumption of intermediates on and off, it can be noted that higher demand for intermediate goods in regions with improved accessibility attracts producers of intermediate goods, which lowers the production costs for the producers of the final goods. In the absence of vertical linkages, the benefits of Cohesion Policy investments are more localized, but, when vertical linkages are allowed, the productivity improvements

[31] See, e.g. Brandsma et al. (2014), which fed the 6th Report on Economic, Social and Territorial Cohesion, European Commission, 2014.

in one region spread to all the regions using its output as an input in their productive processes. Therefore, the benefits of allocating resources to a region are felt beyond its borders.

Such models are powerful tools to check ex ante the potential effects of different transportation policies. However, they suffer from several shortcomings. First, the model is calibrated but not econometrically tested. Second, the mechanisms are so complex, and the model so big, that it is impossible to isolate and identify the drivers of agglomeration and dispersion when all the features are included together. Last, the way the mobility of workers is modeled is critical as European workers are very sticky, while mobility habits can change over time and respond to specific policies (which are impossible to capture accurately in the model). It should also be noted that the administrative capacity of the local authorities and the quality of the planned investments are key determinants of the success of a policy, but these aspects cannot be captured in a general equilibrium approach. To this end, the following approach should complement the ones based on econometric analysis and model simulations.

Assessing transport investments by means of case-studies. In the late 1990s, the EU has selected a priority list of transport investments—the "Trans European Network" investments—whose total value accounted for some €600 billion. These investment projects are the first that should receive European subsidies. In an attempt to assess the benefits of the 22 priority freight projects, Bröcker et al. (2010) have developed a model in the tradition of the new trade theories with 260 European regions. In this model, firms produce a differentiated good and operate under increasing returns and monopolistic competition; interregional trade is costly while capital and labor are immobile. Since production factors are immobile, one major ingredient of NEG is missing, that is, the endogenous formation of clusters. A particular transport investment decreases transport costs between specific regions, which translates into changes in production activities, trade patterns, and ultimately the welfare level of consumers residing in different regions (as in the Cournot model discussed in Section 3.1.1).

There are three main findings for this first round of EU transport priority projects. First, only 12 of the 22 projects pass the cost-benefit analysis test. Second, most projects benefit only the region where the investment takes place, so that the "EU value added"—or the positive spillover argument—does not seem to warrant the investment. Finally, the projects do not systematically favor the poorer regions. Such findings illustrate the role of political economy factors in the selection of projects. According to Knight (2002), the allocation of federal highway funds in the U.S. was highly inefficient in that for every two dollars invested one dollar was wasted. To avoid such a waste of resources, the EU should rely on independent project assessment. There has been great progress in this area over the last decade. The group of countries with a strong tradition of independent project assessment (Netherlands, Sweden, and the UK) has been widened and the methods are being refined to allow for relocation effects.

A second plan of EU transport priority projects has been approved in 2015. The selection of the projects is based on based on expert judgments, which refer to a wide range of objectives, but it is not clear how many projects would pass the CBA test. In total 276 proposals were recommended for funding.[32]

When it comes to passenger transport, the EU has put a strong emphasis on 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, but where HSR projects have never taken off. On average, Americans travel almost 3,000 km per year by air inside the U.S., while the EU citizen travels slightly more than 1,000 km per year by air inside Europe and some 200 km by HSR. Both Americans and Europeans also make long-distance trips by car, but Europeans clearly have lower demands for long-distance trips than Americans.

The EU probably opted for HSR because of the presence of strong (public) national railway companies wanting to preserve their market share. Air transport has grown strongly, and the liberalization of passenger air transport has led to lower prices, higher frequencies, and loss of market share for rail. HSR networks require a large upfront investment in infrastructure (tracks, locomotives). Compared with air transport, HSR has high fixed costs, while infrastructure construction is almost fully subsidized. Maintenance and operation are supposed to be paid for by passenger fares. More investment subsidies are spent on rail than on roads so it is crucial to have a good ex ante appraisal of the different transport modes.

De Rus and Nombela (2007) use standard estimates of stock to determine what the level of demand should be for a HSR link to be socially beneficial. They find that a link needs some 10 million passengers per year and many new HSR links do not meet this target. Adler et al. (2010) use a full-network model where EU passengers have the choice between HSR, air, and car for medium- to long-distance trips. The reactions of the air transport sector are taken into account in order to avoid the mistake made when the Channel Tunnel was assessed without anticipating the reaction of competing ferries. When HSR has to cover all its costs, these authors found 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 right project and cannot serve all regions equally. France and Spain have the largest HSR networks and part of their network would probably not pass the cost-benefit test. The U.K. and the Netherlands have almost no HSR network. Finally, HSR projects are defended by the EU on environmental grounds, but sensitivity analysis shows that one needs extremely high carbon values to make HSR better than air transportation on these grounds.

[32] Recommended projects are very diverse and include a HSR Bordeaux-Dax, a Canal Seine-Escaut, or a Tunnel Lyon – Turin.

3.6.2 Is the EU moving to a better utilization of its existing transport capacity?

Competition on diesel fuel taxes leads EU countries to revise their pricing of road freight. Trucks are responsible for climate damage, conventional air pollution, accidents, and road damage. The main determinant of road damage is the axle weight of a truck. In Europe, trucks pay for the use of roads via excise taxes on diesel fuel but this is changing fast as a result of intense fuel tax competition. Almost all countries charge excise tax on diesel fuel used by trucks. Because trucks can cover from 1,000 to 2,000 km with a single tank of fuel, countries or regions engage in fuel tax competition. The difference in distances covered implies that tax competition is much more important for trucks than for cars. Within the EU, some smaller countries (Luxemburg being the most obvious example) choose a strategy of low diesel excise tax rates to make international haulers fuel up in their country, generating large excise tax revenues for these countries. This strategic behavior has prompted the EU to negotiate a minimum level of excise taxes.

Technological progress in charging techniques implied that several countries with a lot of transit traffic want to introduce distance-based charging. In 2001, Switzerland (not an EU member) replaced its vignette system (a form of road pricing) by a kilometer tax that charges trucks much more than before.[33] The neighboring countries followed: Austria (an alternative route to crossing Switzerland) in 2004, Germany in 2005, the Czech Republic in 2007, Slovakia in 2010, Poland in 2011, and Belgium in 2016.[34]

Replacing diesel taxes by distance charges is not necessarily welfare-improving (Mandell and Proost, 2015). The final outcome depends on the availability of additional instruments to tax diesel cars and on whether the distance charges are finely tuned to external costs or not. To see why, we start with the case where only diesel taxes are used. The level of diesel taxes will depend on the degree of tax competition, which is typically high in international trucking. As shown by Kanbur and Keen (1993), when symmetric countries share a long border, tax competition is tougher and tax rates lower. In contrast, when countries are of different sizes, it pays to be small and to undercut the larger neighbor. In this case, the best strategy for the larger neighbor is not to follow suit.

Mandell and Proost show that, when one country uses distance charges, it can charge all trucks and at the same time undercut the diesel tax of its neighbors. As a consequence, the neighboring countries also have to implement a distance charge if they want to preserve their tax revenues. The end result will be low diesel taxes and high distance charges. Furthermore, when passenger cars also use diesel fuel, taxes are too low for diesel cars while diesel taxes and distance charges are too high for freight transport. Accounting for the inefficient levels of taxes and charges and for the high implementation costs of distance charges, tax competition could lead to a less

[33] Comparing the distance charges across countries shows that Switzerland charges 10 times more per truck-kilometer than most EU countries.

[34] Some EU countries already have a toll system for most of their highways (e.g., France, Italy, and Spain), which serves to cover infrastructure costs.

efficient equilibrium than the fuel tax equilibrium. So the revolution in truck charging, which is a priori an instrument for more efficient pricing, may end up with massive tax exporting.

To some extent, the EU has anticipated that the introduction of distance charges in countries with transit freight traffic may lead to charges that are too high. The EU constitution does not allow discriminatory charges, but this is no guarantee against too-high truck charges in transit countries. It therefore requires distance charges for trucks to be based on external costs. This may be viewed as a principal–agent problem in which the EU is the principal and the country is the agent with better information about external costs. For this reason, distance charges are capped by the EU on the basis of average infrastructure costs. Interestingly, this turns out to be smart policy: when road congestion is an important share of external costs, and road building is governed by constant returns, this cap can guarantee efficient pricing and there is no need for the regulator to know the external congestion costs (Van der Loo and Proost, 2013). The distance charges for trucks up to now have been used chiefly as a simple distance toll with some environmental differentiation. However, the charges can become much more efficient when they are more closely geared to the external costs such as congestion, local air pollution, and accidents. The current revolution in the pricing of trucks may pave the way for a very different charging system for cars.

Finally, we observe that this evolution in the pricing of trucks is largely a European phenomenon. In the U.S., the "stealing" of fuel tax revenues from neighboring states is avoided by a complex system of regularization payments among states, which allows the U.S. to function as an efficient trade zone.

Europe does not make the best use of its rail and air transport system. The EU is still confronted with an archaic rail and air transport system. For rail, there are powerful national regulators and powerful national companies. Rail freight activity has been more or less stable but rail passenger activity has been decreasing substantially over the last 20 years. Rail freight could play a bigger role in freight transport; its market share is 11 percent compared with 40 percent in the U.S. There are probably two reasons for this difference: the lack of consolidation among national companies and the lack of harmonization in operation. Ivaldi and McCullough (2008) studied the integration of freight activities in the U.S. and found that this leads to an important gain in consumer surplus. In the EU, together with the lack of consolidation, there is also a lack of harmonization in the rail business. Harmonization of operating standards is an extremely slow process as the national producers all want to protect their own rail and equipment market.

In the air space, similar mechanisms are at work. In the U.S., there is a single regulator for the management of air space while, in Europe, there are 37 national, and partly regional, monopolies managing air traffic. All regional monopolies function under a cost-plus rule, but an effort is being made to shift to a price-cap system. As a result, costs are almost twice as high as they are in the U.S. Consolidation of different air traffic control zones is doable, which should also lead to important cost reductions. However, it is blocked by the national monopolies.

3.7 What did we learn?

Regional economics, when compared with urban economics, is characterized by a certain lack of unity. However, NEG provides a general set-up in which regional disparities may be analyzed. Saying that NEG is "urban economics without land" is quite a stretch, but it does contain some truth (Fujita and Thisse, 2013). Indeed, the main results presented in this section share some important features with the material surveyed in Section 2. Activities are unevenly distributed across regions, as they are among cities. To a large extent, regional disparities reflect the inequality that cuts across the urban system. In the same vein, large regions may show a high degree of income heterogeneity, just as big and rich cities are often characterized by a strong income and social polarization. And large regions tend to be more attractive than small regions—just as large cities are more attractive than small—but the reasons are not exactly alike. There are two main differences. First, the agglomeration forces are not quite the same because they operate at different spatial scales. Second, the types of spatial friction that affect the distribution of activities are different. Whereas commuting and congestion costs are key in urban economics, transportation costs for goods and interregional mobility costs for workers and capital shape the interregional economy.

Owing to the strength of market forces shaping the spatial economy, regional development seems to be inevitably unequal. Given the First Law of Spatial Economics, not all regions may have a large market populated by skilled workers employed in high-tech industries. The unevenness of regional development may be viewed as the geographical counterpart of economic growth, which is driven mainly by large and innovative cities.[35] The cumulative nature of the agglomeration process makes the resulting pattern of activities particularly robust to various types of shocks, thus showing why it is hard to foster a more balanced pattern of activities. Affluent regions enjoy the existence of agglomeration rents that single-minded policies cannot easily dissipate. Consequently, if the aim of the European Commission is to foster a more balanced distribution of economic activities across European regions, it should add more instruments to its policy portfolio. For example, training people and investing in human capital are often better strategies than building new transport infrastructure, for this heightens the probability of individuals finding a job, maybe in places other than their region of origin. As observed by Cheshire et al. (2014), regional disparities are more driven by differences between workers than by differences between places, although worker and place characteristics interact in subtle ways that require more investigation. After all, Toulouse did not look a priori like a great place for the creation of a top school in economics.

[35] Although care is needed to assess the causality link, it is worth mentioning a recent study by Gennaioli et al. (2013). These authors undertake a cross-sectional analysis of more than 1,500 subnational regions of the world and find that "regional education is the critical determinant of regional development, and the only such determinant that explains a substantial share of regional variation."

A key difficulty highlighted by NEG is that small differences may be enough to trigger regional divergence. This leads to the following question: *when do small differences matter*? As pointed out by Duranton et al. (2010), great places are great "because they have managed to periodically reinvent themselves after losing an important part of the economic fabric." Since the reasons for the success of these cities are often region- or country-specific, it would be futile to seek a universal recipe. Yet, a few general principles may serve as a guide. The historical and social background of a region, its economic strengths and weaknesses, its education system, and its portfolio of amenities are the fundamental ingredients to be accounted for when designing local development policies.

Very much like firms that 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 (Prager and Thisse, 2012). The aim of such strategies is to strengthen regions' comparative advantages and to give priority to finding sustainable solutions to regions' weakest links. For example, by differentiating the infrastructure services they provide, regions can create niches that make them attractive to a certain type of firms, which need not be high-tech firms. The scope for such a strategy is increasing as the revolution in information and communication technology has shifted firms' needs toward more specialized inputs. Implementing such a policy requires precise assessments of the strengths and weaknesses of the regional socio-economic and political fabric. This will be possible only if data collected at various levels (regional, local, household) are made available.

Another unsolved question is the lasting decay that characterizes several regions that used to be the engines of the Industrial Revolution. All industries must one day decline, and examples abound in Europe of old industrialized regions that have succeeded in attracting sizable subsidies to finance inefficient firms. These regions have thus delayed any possibility of the region finding a new niche in which to specialize. Polèse (2010) uses the expression "negative cluster" to describe situations where the (regional) government is captured by a declining cluster that is dominated by a few big employers and trade unions. In addition, it is well documented that the performance of regions in a country also depends on institutions that may be deeply rooted in the past. This leads Polèse (2010) to write: "It is not by accident that the traditional European centres of coal and steel became strongholds of socialist and sometimes communist parties. The era of violent social conflict and divisive labour disputes is today – hopefully – over. But, that era has left a legacy from which some regions have found it more difficult to escape than others. … I can find no other explanation of why seemingly well located regions in northern France and in southern Belgium – in the European heartland – should continue to perform so poorly." This is a strong claim but

part of the story.[36] However, as convincingly argued by Breinlich et al. (2014), we still have a poor understanding of regional decline, which is not the mirror image of regional growth.

In a market economy, the best strategy is to promote efficiency and to complement this policy with a redistribution policy across individuals. Therefore, in a federation like the U.S.—but also in the EU—the core issue is the organization of fiscal federalism, that is, which competencies and fiscal instruments are best centralized and which are best decentralized to lower-tier governments. The game within the EU and the U.S. may be viewed as a sequential game with different leaders and followers. In the U.S., the federal level redistributes income to individuals over the whole territory of the federation and leaves the provision of local public goods to lower-tier governments. As shown by Wildasin (1991) and Wellisch (2000), the scope for state-level redistribution policies is very limited. Because of the high mobility of people, benefits for income transfer recipients (unemployed, retirees, and the like) and the tax rates for these fiscal bases need to be equalized across jurisdictions. Otherwise, migration could be driven by tax or benefit reasons, which would yield an inefficient allocation among regions.

In the EU, however, the roles are reversed. The member countries each take care of redistributing benefits to households. Through such transfers, a rich region A_1 (e.g. Lisbon) of a low-income country A (e.g. Portugal) helps people living in a poor region A_2 of country A, while a rich region B_1 (e.g. Hamburg) of a high-income country B (e.g. Germany) helps the inhabitants of a poor region B_2 of country B. In some cases, region B_2 ends up with a higher income per capita than region A_1. At the subnational level regions can receive EU grants for local public goods that are specified in the cohesion policy objectives. In doing so, the European Commission spends one-third of the EU budget in poor and/or peripheral regions, but the effectiveness of these policies remains an open question (Marzinotto, 2012). Restrictions in the EU redistribution policy (via targeted grants) and asymmetric information between the Commission and member states may result in overspending in particular domains (like transport investments), while reducing incentives for member states and regions to foster economic development. How to design an institutional setting that yields a better redistribution of wealth within the EU should rank high on the research agenda.

Finally, regarding transport investment, there are at least three main research questions that are unsolved. First, given a major transport project, what share of the benefits is triggered by the resulting interregional shift in economic activity—and when does this shift unfold? If it is 10 percent or less, this is within the margin of error of a conventional cost-benefit analysis of a transport project. In contrast, if the share is about 50 percent, a

[36] Shocks that took place long ago may have a lasting influence. For instance, German Protestantism spread from the city of Wittenberg where Martin Luther issued his 95 Theses against indulgences. Becker and Wössman, (2009) show that in 1880 distance from Wittenberg was still an important explanatory factor for the percentage of Protestants in the population, as well as for the level of literacy and industrial development.

conventional cost-benefit analysis is insufficient and must be supplemented by new econometric tools borrowed from urban economics. Second, if small differences in accessibility can have a large impact on the location of economic activity, where is this more likely to happen? Last, how can we make sure that the right transport investments are selected? For example, the EU has been promoting HSR for medium-distance travel, but the selected investments were far from optimal. Another related issue is to make sure that the capacity we have currently is used efficiently.

At present, most interregional road, rail, inland waterways, and air networks are not priced efficiently. Whenever there is a shortage of capacity, peak load pricing that accurately takes into account the different environmental costs is the best instrument to use. It allows one to avoid extending capacity whose benefits are uncertain. But the pricing instrument has not been used. It could be used in the wrong way and the risks of this are easy to understand: as member countries and regions do not take into account the full benefits of international and transit traffic, member countries are incentivized to charge too much for networks used intensively by foreigners. How to avoid this remains a core question. Why pricing is not used more intensively is a political economy or institutional question that has been far too neglected.

3.8 The need for better regional data

Although the challenge of determining the limits of a city will likely remain unmet, defining a region per se is far more problematic (Behrens and Thisse, 2007). Apart from urban regions that are dominated by a large metropolis and have a strong economic identity, many regions are not well-defined economic entities, but often administrative entities (e.g., the NUTS regional classification of the EU). The main challenge lies more in the empirical application one has in mind. Regional studies are often linked to the availability of data. Hence, the question of the spatial scale of analysis, already problematic in theory, becomes even more dramatic in applied research. Furthermore, NUTS-2 regions, say, correspond to areas that have very different economic and physical sizes. For example, Andalucía has a population of 8.4 million inhabitants spread over 87,300 Km2, whereas the Madrid Region accustoms 6.4 million inhabitants concentrated over 8,000 Km2. So one may wonder what we really learn from studies comparing the relative performance of regions that are not really comparable. Furthermore, it is tempting to twist the theory so it fits into the available statistical classifications. Such difficulties should not excuse the analyst from seeking more meaningful solutions. Paul Cheshire's work on "functional urban regions" should be a source of inspiration to others (see Cheshire and Magrini, 2009, for more details).

Data on the cost of living at the regional level are missing in most countries. Interregional comparisons are typically made by using per capita GDP, using a deflator that does not account for housing and commuting costs. Yet these costs vary a lot from one local labor market to another. In a large urban region with a high per capita GDP, housing and commuting costs are much greater than in a region formed by medium-sized and small cities where the per

capita GDP is lower. This leads to overestimating the welfare gaps between regions. For example, Moretti (2013) shows that in the U.S. the real skill premium is significantly smaller that the nominal premium when differences in housing and commuting cost is taken into account.

4. What we know and what we don't know

(i) **The spatial economy is replete with (technological and pecuniary) externalities.** In addition, the nature of externalities changes with the spatial scale. Although externalities are the typical case in which public intervention seems to be desirable, more work is needed to determine which specific policy is needed and at which spatial level it should be implemented. As a first step, avoiding typical mistakes such as those discussed in this survey would already be real progress.

(ii) **The main reason for the existence of cities is to connect people**. 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. Contrary to an opinion widely spread in the media, despite the Internet and other new communication devices, face-to-face contact remains important, at least for certain human and economic activities. Wealth is increasingly created in cities. This holds for the EU and, more generally, for developed and emerging countries alike. What is more, although there is not (yet) an urban strategy at the level of the EU, there is a growing recognition that many large European cities face similar social and cohesion problems. In this respect, the EU cities can be seen as individual experiments from which we can learn, so there is not necessarily a need for harmonized policies.

(iii) **Urban policies are probably more important for economic growth and social cohesion than regional policies**. This is in contrast to the EU's role in designing regional policies and its absence from urban policies. Social tensions between urban neighborhoods are strong and income discrepancies within large cities are wide, and both are growing. Investments in human capital and housing are needed to counter such an evolution, but they will not be sufficient. Several aspects of urban policy suffer from fragmentation of policy areas. This holds for public finance, spatial segregation and housing. Urban transport is characterized by many negative externalities, but the present policy orientations are far from optimal as they not address the most important externality, that is, congestion.

(iv) **Most of the results discussed in this paper suggest a trade-off between global efficiency and spatial equity.** If some cities and regions are richer, it follows that others are less rich or poorer. It thus seems logical, at first sight, to make spatial equity a criterion of economic policy. However, the underlying principles of spatial equity are ambiguous vis-à-vis the principles of social justice. Interpersonal inequality is often larger than interregional inequality, while a redistribution policy based solely on the spatial criterion also benefits all inhabitants of less-developed regions. Helping poor regions does not necessarily mean helping poor people. The poor or unemployed in major urban areas today probably have more right to assistance than the inhabitants of poorer regions with a

substantially lower cost of living. The job of the welfare state is to reduce interpersonal inequalities that run counter to the principles of social justice, and these principles do not refer to particular spatial entities.

(v) **The provision of apparently identical services in all regions may be detrimental to the inhabitants of the regions one wishes to help**. The public services provided may be identical on paper but are in fact of lower quality in poor regions if a critical mass is not attained, as shown by the example of higher education. To some extent, land values reflect the supply of public services, so that the inhabitants of areas with less infrastructure have an advantage that the residents of densely populated areas do not. This difference in land values does not directly influence differences in income, but it does reduce differences in living standards, sometimes substantially.

(vi) **How does a lowering of interregional transport costs affect the location of activity?** One would expect the market-access effect to be weaker when transport costs are lower. The opposite often holds true: more firms choose to set up in the large markets when it gets cheaper to trade goods between regions. Lower transport costs makes exports to the small markets easier, which allows firms to exploit their scale economies more intensively by locating in in the large ones. But lower transport costs also reduce the advantages associated with geographical isolation in the small markers where there is less competition. These two effects push toward more agglomeration. Hence, connecting lagging regions to dynamic urban centers may weaken their industrial base. This result may come as a surprise to those who forget that highways run both ways. .

(vii) **More attention should be given to the quality of local institutions**. Even with the best will in the world, outsiders cannot generate high regional growth or make sure a city tackles its problems successfully: the final outcome depends on the ability of the local institutions to implement the policies and reforms needed to boost economic development (Glaeser and Shleifer, 2001; Prager and Thisse, 2012). There is room for public policies, but they must be tested and assessed carefully. Many of them have been disappointing so far, which undermines the credibility of the polity. The coming Junker Investment Plan (€315 billion) may provide an important economic stimulus in the short, and perhaps medium, run but the selection of projects may be a mixed bag for global efficiency and spatial inequality within the EU. The EU is replete with examples where well-meaning policies that were adopted were in the end worse than having no policy at all – to quote Francis Bacon: "*The remedy was worse than the disease*". The minimum goal of institutions involved in regional development must at least be to ensure that they avoid that state of affairs.

References

Adler, N., C. Nash and E. Pels (2010) High-speed rail and air transport competition: Game engineering as tool for cost-benefit analysis. *Transportation Research Part B* 44: 812-33.

Ahlfeldt, G.M. and A. Feddersen (2015) From periphery to core: Measuring agglomeration effects using high-speed rail. SERC Discussion Papers 0172, Spatial Economics Research Centre, LSE

Alonso, W. (1964) *Location and Land Use*. Cambridge, MA: Harvard University Press.

Albouy, D. (2009) The unequal geographic burden of federal taxation**.** *Journal of Political Economy* 117: 635-68.
Anas, K. and I. Kim (1996) General equilibrium models of polycentric urban land use with endogenous congestion and job agglomeration. *Journal of Urban Economics* 40**:** 232-56.
Anas, A. and R. Lindsey (2011) Reducing urban road transportation externalities: Road pricing in theory and in practice. *Review of Environmental Economics and Policy* 5: 66-88.
Anas, A. and Y. Liu (2007) A regional economy, land use, and transportation model (RELU-TRAN): Formulation, algorithm design, and testing. *Journal of Regional Science* 47: 415-55.
Anderson, J. and E. van Wincoop (2004) Trade costs. *Journal of Economic Literature* 42: 691-751.
Arnott, R. (1979) Unpriced transport congestion. *Journal of Economic Theory* 21: 294-316.
Arnott, R. (1981) Aggregate land rents and aggregate transport costs. *Economic Journal* 91: 331-47.
Arnott, R. (2004) Does the Henry George Theorem provide a practical guide to optimal city size? *American Journal of Economics and Sociology* 63: 1057-90.
Arnott, R. (2007) Congestion tolling with agglomeration externalities. *Journal of Urban Economics* 62: 187-203.
Arnott, R., A. de Palma and R. Lindsey (1993) A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *American Economic Review* 83: 161-79.
Arnott, R. and K. Small (1994) The economics of traffic congestion. *American Scientist* 82: 446-55.
Arzaghi, M. and J.V. Henderson (2008) Networking off Madison Avenue. *Review of Economic Studies* 75: 1011-38.
Aschauer, D.A. (1989) Is public expenditure productive? *Journal of Monetary Economics* 23; 177-200.
Au, C.C. and J.V. Henderson (2006) Are Chinese cities too small? *Review of Economic Studies* 73: 549-76.
Bacolod, M., B.S. Bull and W.C. Strange (2009) Skills in the city. *Journal of Urban Economics* 65: 136-53.
Bairoch, P. (1985) *De Jéricho à Mexico. Villes et économie dans l'histoire*. Paris: Editions Gallimard. English translation: *Cities and Economic Development: From the Dawn of History to the Present*. Chicago: The University of Chicago Press, 1988.
Bairoch, P. (1997) *Victoires et déboires. Histoire économique et sociale du monde du XVIe siècle à nos jours*. Paris: Editions Gallimard.
Baldwin, R.E. and P.R. Krugman (2004) Agglomeration, integration and tax harmonization. *European Economic Review* 48: 1-23.
Baldwin, R.E., R. Forslid, P. Martin, G.I.P. Ottaviano and F. Robert-Nicoud (2003) *Economic Geography and Public Policy*. Princeton, NJ: Princeton University Press.
Barbero, J., K. Behrens and J.L. Zofio (2015) Industry location and wages: The role of market size and accessibility in trading networks. CEPR Discussion Paper N°10411.
Barnett, J. (1986) *The Elusive City. Five Centuries of Design, Ambition and Miscalculation*. New York: Harper and Row.
Basso L.J. and H.E. Silva (2014) Efficiency and substitutability of transit subsidies and other urban transport policies. *American Economic Journal: Economic Policy* 6: 1-33.
Barrios, S. and E. Strobl (2009) The dynamics of regional inequalities. *Regional Science and Urban Economics* 39: 575–91.
Batty, M. (2013) *The New Science of Cities*. Cambridge, MA: MIT Press.
Baum-Snow, N. and F. Ferreira (2015) Causal inference in urban and regional economics. In: G. Duranton, J.V. Henderson and W. Strange (eds.) *Handbook of Regional and Urban Economics. Volume 5*. Amsterdam: Elsevier, 3–68.
Becker, S. and L. Wößman (2009) Was Weber wrong? A human capital theory of protestant economic history. *Quarterly Journal of Economics* 124: 531–96.
Beckmann, M.J. (1976) Spatial equilibrium in the dispersed city. In: Y.Y. Papageorgiou (ed.). *Mathematical Land Use Theory*. Lexington, MA: Lexington Books, 117-25.
Beckmann, M.J. and J.-F. Thisse (1986) The location of production activities. In: P. Nijkamp (ed.). *Handbook of Regional and Urban Economics, Volume 1*. Amsterdam: Elsevier, 21-95.

Belleflamme, P., P. Picard and J.-F. Thisse (2000) An economic theory of regional clusters. *Journal of Urban Economics* 48: 158-84.

Behrens, K. (2005) Market size and industry location: Traded vs non-traded goods. *Journal of Urban Economics* 58: 24-44.

Behrens, K. and F. Robert-Nicoud (2015) Agglomeration theory with heterogeneous agents. In: G. Duranton, J.V. Henderson and W. Strange (eds.) *Handbook of Regional and Urban Economics. Volume 5*. Amsterdam: Elsevier, 171–245.

Behrens, K. and J.-F. Thisse (2007) Regional economics: A new economic geography perspective. *Regional Science and Urban Economics* 37: 457-65.

Behrens, K., G. Duranton and F. Robert-Nicoud (2014) Productive cities: Sorting, selection, and agglomeration. *Journal of Political Economy* 122: 507-53.

Behrens, K., C. Gaigné, G.I.P. Ottaviano and J.-F. Thisse (2007) Countries, regions and trade: On the welfare impacts of economic integration. *European Economic Review* 51: 1277-301.

Behrens, K., A.R. Lamorgese, G.I.P. Ottaviano and T. Tabuchi (2009) Beyond the home market effect: Market size and specialization in a multi-country world. *Journal of International Economics* 79: 259-65.

Bénabou, R. (1993) Workings of a city: Location, education, and production. *Quarterly Journal of Economics* 108: 619-52.

Bernard, A., J. Jensen, S. Redding and P. Schott (2007) Firms in international trade. *Journal of Economic Perspectives* 21/3: 105-30.

Bloom, N., J. Liang, J. Roberts and Z.-J Ying (2015) Does working from home work? Evidence from a Chinese experiment. *Quarterly Journal of Economics* 130: 165-218.

Boldrin, M. and F. Canova (2001) Inequality and convergence in Europe¹s regions: Reconsidering European regional policies. *Economic Policy* 16: 207-53.

Bonfatti, R. and S. Poelhekke (2015) From mine to coast: Transport infrastructure and the direction of trade in developing countries. Tinbergen Institute, mimeo.

Bosquet, C. and H. Overman (2015) Home versus hometown: What do we mean by spatial sorting? Paper presented at the RIETI workshop "Frontiers in Spatial Economics," Tokyo, April 14, 2015.

Brakman, S. and C. van Marrewijk (2008) It's a big world after all: On the economic impact of location and distance. *Cambridge Journal of Regions, Economy and Society* 1: 411-37.

Brandsma, A., F. Di Comite, O. Diukanova, d'A. Kancs, J. López-Rodríguez, D. Persyn and L. Potters (2014) Assessing policy options for the EU Cohesion Policy 2014-2020. *Investigaciones regionales* 29: 17- 46.

Breinlich, H., G. Ottaviano and J. Temple (2014) Regional growth and regional decline. In: P. Aghion and S.N. Durlauf (eds.) *Handbook of Economic Growth. Volume 2*. Amsterdam: Elsevier, 683-779.

Briant, A., M. Lafourcade and B. Schmutz (2015) Can tax breaks beat geography? Lessons from the French enterprise zone experience. *American Economic Journal: Economic Policy* 7: 88-124.

Bröcker, J., A. Korhenevych and C. Schürmann (2010) Assessing spatial equity and efficiency impacts of transport infrastructure projects. *Transportation Research Part B* 44: 795-811.

Bröcker, J. and J. Mercenier (2011) General equilibrium models for transportation economics. In: A. de Palma, R. Lindsey, E. Quinet, and V. Vickerman (eds.) *Handbook in Transport Economics*. Cheltenham, UK: Edward Elgar, 21–45.

Brueckner J.F. and S. F. Franco (2015) Parking and urban form. University of California at Irvine, memo.

Brueckner, J. K. and S.K. Sridhar (2012) Measuring welfare gains from relaxation of land-use restrictions: The case of India's building-height limits. *Regional Science and Urban Economics* 42: 1061-67.

Brueckner, J.K., J.-F. Thisse and Y. Zenou (1999) Why is Central Paris rich and Downtown Detroit poor? An amenity-based theory. *European Economic Review* 43: 91-107.

Brülhart, M., S. Bucovetsky and K. Schmidheiny (2015) Taxes in cities. In: G. Duranton, J.V. Henderson and W. Strange (eds.) *Handbook of Regional and Urban Economics. Volume 5*. Amsterdam: Elsevier, 1123-96.

Cervero, R. (2003) Road expansion, urban growth, and induced travel: A path analysis. *Journal of the American Planning Association* 69: 145–63.

Charlot, S. and G. Duranton (2004) Communication externalities in cities. *Journal of Urban Economics* 56: 581-613.

Charlot, S., C. Gaigné, F. Robert-Nicoud and J.-F. Thisse (2006) Agglomeration and welfare: the core-periphery model in the light of Bentham, Kaldor, and Rawls. *Journal of Public Economics* 90: 325-47.

Cheshire, P. and S. Magrini (2006) Population growth in European cities: Weather matters -- but only nationally. *Regional Studies* 40: 23-37.

Cheshire, P. and S. Magrini (2009) Urban growth drives in a Europe of sticky people and implicit boundaries. *Journal of Economic Geography* 9: 85-116.

Cheshire, P., M. Nathan and H. Overman (2014) *Urban Economics and Urban Policy*. Cheltenham, U.K.: Edward Elgar.

Ciccone, A., and R.E. Hall (1996) Productivity and the density of economic activity. *American Economic Review 8*6: 54-70.

Combes, P.-P., G. Duranton and L. Gobillon (2008) Spatial wage disparities: Sorting matters! *Journal of Urban Economics* 63: 723-42.

Combes, P.-P., G. Duranton and L. Gobillon (2011) The identification of agglomeration economies. *Journal of Economic Geography* 11: 253-66.

Combes, P.-P., G. Duranton, L. Gobillon, D. Puga and S. Roux (2012) The productivity advantages of large cities: Distinguishing agglomeration from firm selection. *Econometrica* 80: 2543-25.

Combes, P.-P. and L. Gobillon (2015) The empirics of agglomeration economies. In: G. Duranton, J.V. Henderson and W. Strange (eds.) *Handbook of Regional and Urban Economics. Volume 5*. Amsterdam: Elsevier, 247-348.

Combes, P.-P. and M. Lafourcade (2005) Transport costs: Measures, determinants, and regional policy implications for France. *Journal of Economic Geography* 5: 319-49.

Combes, P.-P., M. Lafourcade, J.-F. Thisse and J.C. Toutain, (2011) The rise and fall of spatial inequalities in France. A long-run perspective. *Exploration in Economic History* 48: 243-71.

Combes, P.-P. and H. Overman (2004) The spatial distribution of economic activities in the European Union. In: J.V. Henderson and J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics. Cities and Geography*. Amsterdam: Elsevier, 2845–909.

Cournot, A. (1838) *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette. English translation: *Researches into the Mathematical Principles of the Theory of Wealth*. New York: Macmillan (1897).

Crozet, M. (2004) Do migrants follow market potentials? An estimation of a new economic geography model. *Journal of Economic Geography* 4: 439-58.

d'Aspremont, C., Gabszewicz, J.J., and Thisse, J.-F. (1979) On Hotelling's "Stability in Competition". *Econometrica* 47: 1045-50.

D'Haultfoeuille X., P. Givord and X. Boutin (2013) The environmental effect of green taxation: The case of the French "Bonus/Malus". *Economic Journal* 124: 444-80.

Davis, M. and F. Ortalo-Magné (2011) Household expenditures, wages, rents. *Review of Economic Dynamics* 14: 248-61.

Davis, D.R. and D.E. Weinstein (2002) Bones, bombs, and break points: The geography of economic activity. *American Economic Review* 92: 1269-89.

Davis, D. R. and D.E. Weinstein (2003) Market access. Economic geography and comparative advantage: An empirical assessment. *Journal of International Economics* 59: 1-23.

Diamond, R. (2015) The determinants and welfare implications of US workers diverging location choices by skill: 1980-2000. Stanford University, memo.

De Borger, B. and S. Proost (2012) A political economy model of road pricing. *Journal of Urban Economics* 71: 79-92.

de Palma, A., M. Bierlair, R. Hurtibia and P. Waddell (2015) Future challenges in transport and land use planning. In: M. Bierlaire, A. de Palma, R. Hurtibia and P. Waddell (eds.) *Integrated Transport and Land Use Planning*. Lausanne: EPFL Press.

De Rus, G. and G. Nombela (2007) Is investment in high speed rail socially profitable? *Journal of Transport Economics and Policy* 41: 3-23.

Desmet, K. and E. Rossi-Hansberg (2013) Urban accounting and welfare. *American Economic Review* 103: 2296-327.

Di Comite, F. and A. Kancs (2014) Modelling agglomeration and dispersion in RHOMOLO. JRC Technical Report.

Diewert, E. and C. Shimizu (2015) Residential property price indexes for Tokyo. *Macroeconomic Dynamics*, forthcoming.

Dixit, A.K. and J.E. Stiglitz (1977) Monopolistic competition and optimum product diversity. *American Economic Review* 67: 297-308.

Duranton, G., and D. Puga (2001) Nursery cities: Urban diversity, process innovation, and the life-cycle of products. *American Economic Review* 91: 1454-63.

Duranton, G. and D. Puga (2004) Micro-foundations of urban increasing returns: Theory. In: J.V. Henderson and J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics. Cities and Geography*. Amsterdam: Elsevier, 2063-117.

Duranton, G. and M.A. Turner (2011) The fundamental law of road congestion: Evidence from US cities. *American Economic Review* 101: 2616-52.

Duranton, G., P. Martin, T. Mayer and F. Mayneris (2010) *The Economics of Clusters*. Oxford: Oxford University Press.

Eaton, J. and Z. Eckstein (1997) Cities and growth: Theory and evidence from France and Japan. *Regional Science and Urban Economics* 27: 443–74.

Eeckhout, J. R. Pinheiro and K. Schmidheiny (2014) Spatial sorting. *Journal of Political Economy* 122:European Commission (2014) *Handbook on External Costs*. DG-Move, report Ricardo-AEA/R/ ED57769.

Eliasson, J., L. Hultzkranz, L. Nerhagen and L. Smidfelt Rosqvist (2009) The Stockholm congestion charging trial 2006: Overview of effects. *Transportation Research A* 43: 240-50.

Eliasson J. and S. Proost (2015) Is sustainable transport sustainable? *Transport Policy* 37: 92-100.

Ellison, G., E.L. Glaeser and W.R. Kerr (2010) What causes industry agglomeration? Evidence from coagglomeration patterns. *American Economic Review* 100: 1195-213.

European Commission (2015) Connecting Europe Facility – proposal for the selection of projects, Directorate General for Mobility and Transport.

Faber, B. (2014) Trade integration, market size, and industrialization: Evidence from China's national trunk highway system. *Review of Economic Studies* 81: 1046-10.

Faggio, G., O. Silva and W.C. Strange (2015) Heterogeneous agglomeration. *Review of Economics and Statistics*, forthcoming.

Faini, R. (1999) Trade union and regional development. *European Economic Review* 43: 457-74.

Falck, O., S. Heblich, A. Lameli and J. Südekum (2012) Dialects, cultural identity, and economic exchange. *Journal of Urban Economics* 72: 225-39.

Figueiredo, O, P. Guimaraes and D. Woodward (2014) Firm--worker matching in industrial clusters. *Journal of Economic Geography* 14: 1-19.

Feenstra, R. C. (1998) Integration of trade and disintegration of production in the global economy. *Journal of Economic Perspectives* 12 (Fall): 31-50.

Fogelson, R.M., 2005. *Downtown. Its rise and fall, 1880-1950*. New Haven: Yale University Press.

Frank, B. (2014) Laboratory evidence on face-to-face: Why experimental economics is of interest to regional economists. *International Regional Science Review* 37: 411-35.

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

Fujita, M., P. Krugman and A.J. Venables (1999) *The Spatial Economy. Cities, Regions and International Trade*.

Cambridge, MA: MIT Press.
Fujita, M. and J.-F. Thisse (2006) Globalization and the evolution of the supply chain: Who gains and who loses? *International Economic Review* 47: 811-36.
Fujita, M. and J.-F. Thisse (2013) *Economics of Agglomeration. Cities, Industrial Location and Globalization. Second edition*. Cambridge, Cambridge University Press.
Gabaix, X. and Y.M. Ioannides (2004) The evolution of city size distributions. In: J.V. Henderson and J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics. Cities and Geography*. Amsterdam: Elsevier, 2341-78.
Gagnepain, P., M. Ivaldi and D. Martimort (2013) The cost of contract renegotiation: Evidence from the local public sector. *American Economic Review* 103: 2352-83.
Gaigné, C., S. Riou and J.-F. Thisse (2013) How to make the metropolitan area work? Neither big government, nor laissez-faire. CEPR Discussion Paper N° 9499.
García-López, M.A., A. Holl and E. 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.
Gaspar, J. and E.L. Glaeser (1998) Information technology and the future of cities. *Journal of Urban Economics* 43: 136-56.
Gennaioli, N., R. La Porta, F. Lopez-de-Silanes and A. Shleifer (2013) Human capital and regional development. *Quarterly Journal of Economics* 128: 105 - 64.
Gibbons, S., H.G. Overman and P. Pelkonen (2014) Area disparities in Britain: Understanding the contribution of people vs. place through variance decompositions. *Oxford Bulletin of Economics and Statistics* 76: 745-63.
Glaeser, E.L. (2011) *Triumph of the City*. London: Macmillan.
Glaeser, E. and J. Gyourko (2003) The impact of building restrictions on housing affordability. *FRBNY Economic Policy Review*, 9, pp. 21-39.
Glaeser, E.L. and M.E. Kahn (2004) Sprawl and urban growth. In: J.V. Henderson and J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics. Cities and Geography*. Amsterdam: Elsevier, 2481-527.
Glaeser, E.L. and J.E. Kohlhase (2004) Cities, regions and the decline of transport costs. *Papers in Regional Science* 83: 197-228.
Glaeser, E.L., M.E. Kahn and J. Rappaport (2008) Why do the poor live in cities? The role of public transportation. *Journal of Urban Economics* 63: 1-24.
Glaeser, E.L., H.D. Kallal, J.A. Scheinkman and A. Shleifer (1992) Growth in cities. *Journal of Political Economy* 100: 1126-52.
Glaeser, E.L., J. Kolko and A. Saiz (2001) Consumer city. *Journal of Economic Geography* 1: 27-50.
Glaeser, E.L. and D. Maré (2001) Cities and skills. *Journal of Labor Economics* 19: 316-42.
Glaeser, E. and A. Shleifer (2001) Not-for-profit entrepreneurs. *Journal of Public Economics* 81: 99-115.
Gramlich, E.M. (1994) Infrastructure investment: A review essay. *Journal of Economic Literature* 32: 1176-96.
Gobillon, L. and C. Milcent (2013) Spatial disparities in hospital performance. *Journal of Economic Geography* 13: 1013-40.
Grigolon L., Reynaert, M. and F. Verboven, (2014) Consumer valuation of fuel costs and the effectiveness of tax policy: Evidence from the European car market. *CES Discussion Paper* 14.34. KULeuven.
Groot, S.P.T., H.L.F. Groot and M.J. Smit (2014) Regional wage differences in the Netherlands: Micro evidence on agglomeration externalities. *Journal of Regional Science* 54: 503-23.
Guo, Z. and Ren, S. (2013) From minimum to maximum: impact of the London parking reform on residential parking supply from 2004 to 2007. *Urban Studies* 52: 1183-200.
Handbury, J. and D. Weinstein (2015) Goods prices and availability in cities. *Review of Economic Studies* 82: 258-96.
Harding, M. (2014) The personal tax treatment of company cars and commuting expenses: Estimating the Fiscal and Environmental costs. *OECD Taxation working papers* n°20, OECD Publishing, Paris.
Harris, C. (1954) The market as a factor on the localization of industry in the United States. *Annals of the Association of American Geographers* 64: 315-480.

Head, K. and T. Mayer (2004) The empirics of agglomeration and trade. In: J.V. Henderson and J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics. Volume 4*. Amsterdam: Elsevier, 2609-69.
Head, K. and T. Mayer (2011) Gravity, market potential and economic development. Journal of Economic Geography 11: 281--94.
Head, K. and T. Mayer (2014) Gravity equations: Workhorse, toolkit, and cookbook. In: G. Gopinath, E. Helpman and K. Rogoff (eds.) *Handbook of International Economics, Volume 4*. Amsterdam: Elsevier.
Helpman, E. (1998) The size of regions. In: D. Pines, E. Sadka and I. Zilcha (eds.). *Topics in Public Economics. Theoretical and Applied Analysis.* Cambridge: Cambridge University Press, 33-54.
Helsley, R.W. and W.C. Strange (2014) Co-agglomeration and the scale and composition of clusters. *Journal of Political Economy* 122: 1064-93.
Henderson, J.V. (1974) The sizes and types of cities. *American Economic Review* 64: 640-56.
Henderson, J.V. (1988) *Urban Development. Theory, Fact and Illusion*. Oxford: Oxford University Press.
Henderson, J.V. (1997) Medium size cities. *Regional Science and Urban Economics* 27: 583-612.
Henderson, V., A. Kuncoro and M. Turner (1995) Industrial development in cities. *Journal of Political Economy* 103: 1066-90.
Hillberry, R. and D. Hummels (2007) Trade responses to geographic frictions: A decomposition using micro-data. *European Economic Review* 52: 527-50.
Hirschman, A.O. (1958). *The Strategy of Development*. New Haven, CN: Yale University Press.
Hochman, O., D. Pines and J.-F. Thisse (1995) On the optimal structure of local governments. *American Economic Review* 85: 1224-40.
Hohenberg, P. and L.H. Lees (1985) *The Making of Urban Europe (1000-1950)*. Cambridge, MA: Harvard University Press.
Hotelling, H. (1929). Stability in competition. *Economic Journal* 39: 41-57.
ICCT (2013) EU - vehicle market pocket book, International Council on Clean Transportation, http://eupocketbook.theicct.org
Inci, E. (2015) A review of the economics of parking, *Economics of Transportation*, forthcoming
Ioannides, Y.M. (2012) *From Neighborhoods to Nations: The Economics of Social Interactions*. Princeton, NJ: Princeton University Press.
Jacobs, J. (1969) *The Economy of Cities*. New York: Random House.
Jackson, M.O. (2008) *Social and Economic Networks*. Princeton, NJ: Princeton University Press.
Jofre-Monseny, J. (2013) Is agglomeration taxable? *Journal of Economic Geography* 13: 177-201.
Kanbur, R. and M. Keen (1993) Jeux sans frontières: Tax competition and tax coordination when countries differ in size, *American Economic Review* 83: 877-92.
Klein, A. and N. Crafts (2012) Making sense of the manufacturing belt: Determinants of U.S. industrial location, 1880-1920. *Journal of Economic Geography* 12: 775-807.
Knight B. (2002) Parochial interests and the centralized provision of local public goods: Evidence from congressional voting on transportation projects. *Journal of Public Economics* 88: 845-66.
Koenig, P., F. Mayneris and S. Poncet (2010) Local export spillovers in France. *European Economic Review* 54: 622-41.
Koh, H.-J., N. Riedel and T. Böhm (2013) Do governments tax agglomeration rents? *Journal of Urban Economics* 75: 92-106.
Koopmans, T.C. (1957) *Three Essays on the State of Economic Science*. New York: McGraw-Hill.
Krugman, P.R. (1980) Scale economies, product differentiation, and the pattern of trade. *American Economic Review* 70: 950-959.
Krugman, P. (1991) Increasing returns and economic geography. *Journal of Political Economy* 99: 483-99.
Krugman, P. and A.J. Venables (1995) Globalization and the inequality of nations. *Quarterly Journal of Economics* 110: 857-80.
Launhardt, W. (1885) *Mathematische Begründung der Volkswirtschaftslehre*. Leipzig: B.G. Teubner. English

translation: *Mathematical Principles of Economics*. Aldershot: Edward Elgar (1993).
Leamer, E.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.
Lösch, A. (1940) *Die Räumliche Ordnung der Wirtschaft*. Jena: Gustav Fischer. English translation: *The Economics of Location*. New Haven, CN: Yale University Press (1954).
Lowry, I.S. (1964) A model of metropolis. Memorandum RM-4035-RC. Santa Monica, CA: The Rand Corporation.
Lucas, R.E. (1988) On the mechanics of economic development. *Journal of Monetary Economics* 22: 3-42.
Luthi, E. and K. Schmidheiny (2014) The effect of agglomeration size on local taxes. *Journal of Economic Geography* 14: 265-87.
Magrini, S. (2004) Regional (di)convergence. In: J.V. Henderson and J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics. Cities and Geography*. Amsterdam: Elsevier, 2741-96.
Mandell, S. and S. Proost (2015) Why truck distance taxes are contagious and drive fuel taxes to the bottom. KULeuven, CES-Discussion Paper N° 15.04
Marshall, A. (1890) *Principles of Economics*. London: Macmillan. 8th edition published in 1920.
Marzinotto, B. (2012) The growth effects of EU cohesion policy: A meta-analysis. Breugel Working Paper N° 2012/14.
Mayer, T. (2008) Market potential and development. CEPR Discussion Paper N°6798.
Mayer, T. and G.I.P. Ottaviano (2007) *The Happy Few: The Internationalisation of European Firms*. Brussels: Bruegel Blueprint Series.
Mayeres, I. and S. Proost (2001) Marginal tax reform, externalities and income distribution. *Journal of Public Economics* 79: 343-63.
Melo, P., D. Graham and R. Brage-Ardao (2013) The productivity of transport infrastructure investment: A meta-analysis of empirical evidence. *Regional Science and Urban Economics* 43: 695-706.
Midelfart-Knarvik, K. H. and H. G. Overman (2002) Delocation and European Integration: Is structural spending justified? *Economic Policy* 17: 321-60.
Mills, E.S. (1967) An aggregative model of resource allocation in a metropolitan area. *American Economic Review* 57: 197-210.
Mion, G. and P. Naticchioni (2009) The spatial sorting and matching of skills and firms. *Canadian Journal of Economics* 42: 28-55.
Mohring, H. (1972) Optimization and scale economies in urban bus transportation. *American Economic Review* 62: 591–604.
Moretti, E. (2004) Estimating the external returns to higher education: Evidence longitudinal and repeated cross-sectional data. *Journal of Econometrics* 121: 175-212.
Moretti, E. (2011) Local labor markets. In: D. Card and O. Ashenfelter (eds.) *Handbook of Labor Economics. Volume 4, Part B*. Amsterdam: Elsevier, 1237-313.
Moretti, E. (2012) *The New Geography of Jobs*. Boston: Houghton Mifflin Harcourt.
Moretti, E. (2013) Real wage inequality. *American Economic Journal: Applied Economics* 5: 65-103,
Munk-Nielsen, A. (2014) Diesel cars and environmental policy. Paper presented at the Conference on the future of fuel taxes, CTS-KTH Stockholm, September 2014.
Muth, R.F. (1969) *Cities and Housing*. Chicago: University of Chicago Press.
Nitsch, V. (2005) Zipf zipped. *Journal of Urban Economics* 57: 86–100.
Nocke, V. (2006) A gap for me: Entrepreneurs and entry. *Journal of the European Economic Association* 4: 929-55.
Ogawa, H. and M. Fujita (1980) Equilibrium land use patterns in a non-monocentric city. *Journal of Regional Science* 20: 455-75.
Ohlin, B. (1933) *Interregional and International Trade.* Cambridge, MA: Harvard University Press). Revised version published in 1968.
Okubo, T., P. Picard, and J.-F. Thisse (2010) The spatial selection of heterogeneous firms. *Journal of International Economics* 82: 230-37.

O'Rourke, K.H. and J.G. Williamson (1999) *Globalization and History*. *The Evolution of a Nineteenth Century Atlantic Economy*. Cambridge, MA: The MIT Press.

Ota, M. and M. Fujita (1993) Communication technologies and spatial organization of multi-unit firms in metropolitan areas. *Regional Science and Urban Economics* 23: 695-729.

Ottaviano G.I.P., T. Tabuchi and J.-F. Thisse (2002) Agglomeration and trade revisited. *International Economic Review* 43: 409-36.

Ottaviano G. and J.-F. Thisse (2002) Integration, agglomeration and the political economics of factor mobility. *Journal of Public Economics* 83: 429-56.

Ottaviano, G.I.P., and T. van Ypersele (2005) Market size and tax competition. *Journal of International Economics* 67: 25-46.

Parry, I. and A. Bento (2002) Revenue recycling and the welfare effects of road pricing. *Scandinavian Journal of economics* 103: 645-71.

Parry, I. and K.A. Small (2005) Does Britain or the United States have the right gasoline tax. *American Economic Review* 95: 1276-89.

Parry, I.W.H. and K.A. Small (2009) Should urban subsidies be reduced? *American Economic Review* 99: 700-24.

Parry, I.W.H., D. Evans and W.E. Oates (2014) Are energy efficiency standards justified? *Journal of Environmental Economics and Management* 67: 104-25.

Pflüger, M. and S. Südekum (2008) Integration, agglomeration, and welfare. *Journal of Urban Economics* 63: 544-66.

Picard, P. and T. Okubo (2012) Firms' location under demand heterogeneity. *Regional Science and Urban Economics* 42: 961-74.

Pigou, A.C. (1920) *The Economics of Welfare*. 4th ed. London: Macmillan. Available online at: http://www.econlib.org/library/NPDBooks/Pigou/pgEW.html

Pirenne, H. (1925) *Medieval Cities*. Princeton: Princeton University Press.

Polèse, M. (2010) *The Wealth and Poverty of Regions. Why Cities Matter*. Chicago: University of Chicago Press.

Pollard, S. (1981) *Peaceful Conquest: The Industrialization of Europe 1760-1970*. Oxford: Oxford University Press.

Porter, M. (1998) Clusters and the new economics of competition. *Harvard Business Review* (November-December): 77-90.

Prager, J.-C. and J.-F. Thisse (2012) *Economic Geography and the Unequal Development of Regions*. London: Routledge.

Proost, S. and K. Van Dender (2012) Energy and environment challenges in the transport sector. *Economics of Transportation* 1: 77-87.

Puga, D. (1999) The rise and fall of regional inequalities. *European Economic Review:* 303–34.

Puga, D. (2002) European regional policies in the light of recent location theories. *Journal of Economic Geography* 2: 373-406.

Puga, D. (2010) The magnitude and causes of agglomeration economies. *Journal of Regional Science* 50: 203-19.

Redding, S. (2011) Economic geography: A review of the theoretical and empirical literature. In: D. Bernhofen, R. Falvey, D. Greenaway and U. Kreickemeie (eds.) *The Palgrave Handbook of International Trade*. London: Palgrave Macmillan.

Redding, S. and D. Sturm (2008) The cost of remoteness: Evidence from German division and reunification. *American Economic Review* 98: 1766–97.

Quinet E. and A. Raj (2015) Welfare and growth effects in transport and trade – a practioner's point of view. Report prepared for the World Bank.

Redding, S.J. and M.A. Turner (2015) Transportation costs and the spatial organization of economic activity. In: G. Duranton, J.V. Henderson and W. Strange (eds.) *Handbook of Regional and Urban Economics. Volume 5*, Amsterdam: Elsevier, 1339-98.

Redding, S.J. and A.J. Venables (2004) Economic geography and international inequality. *Journal of International Economics* 62: 53-82.

Robert-Nicoud, F. (2005) The structure of simple `New Economic Geography' models (or, On identical twins). *Journal of Economic Geography* 5: 201-34.

Rodriguez-Pose, A. and M. Vilalta-Bufia (2005) Education, migration, and job satisfaction: The regional returns of human capital in the EU. *Journal of Economic Geography* 5: 545–66.

Romer, P. (1990) Endogenous technological change. *Journal of Political Economy* 98: S71-S102.

Rosenthal, S.S. and W.C. Strange (2004) Evidence on the nature and sources of agglomeration economies. In: J.V. Henderson and J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics. Cities and Geography*. Amsterdam: Elsevier, 2119-71.

Samuelson, P.A. (1952) Spatial price equilibrium and linear programming. *American Economic Review* 42: 283-303.

Samuelson, P.A. (1954) The transfer problem and transport cost, II: Analysis of effects of trade impediments. *Economic Journal* 64: 264-89.

Sandmo, A. (2011) *Economics Evolving. A History of Economic Thought*. Princeton, NJ: Princeton University Press.

Schelling, T.C. (1971) Dynamic models of segregation. *Journal of Mathematical Sociology* 1: 143–86.

Schmidheiny, K and J. Südekum (2015) The pan-European population distribution across consistently defined functional urban areas. *Economics Letters* 133: 10–13.

Small K.A. and E.T. Verhoef (2007) *Economics of Urban Transportation*. London: Routledge.

Solé-Ollé, A. and E. Viladecans-Marsal (2012) Lobbying, political competition and local land supply: recent evidence from Spain. *Journal of Public Economics* 96: 10-19.

Solow, R.M. (1973) On equilibrium models of urban locations. In: J.M. Parkin (ed.). *Essays in Modern Economics*. London: Longman, 2-16.

Spulber, D. F. (2007) *Global Competitive Strategy*. Cambridge: Cambridge University Press.

Stiglitz, J. (1977) The theory of local public goods. In: M.S. Feldstein and R.P. Inman (eds.) *The Economics of Public Services*. London: Macmillan, 273-334.

Storper, M. (2013) *Keys to the City*. Princeton: Princeton University Press.

Syverson, C. (2004) Market structure and productivity: A concrete example. *Journal of Political Economy* 112: 1181-222.

Tabuchi, T., J.-F. Thisse and X. Zhu (2015) Does technological progress affect the location of economic activity? CORE Discussion Paper N°2015/11.

Takahashi, T., H. Takatsuka and D.-Z. Zeng (2013) Spatial inequality, globalization, and footloose capital. *Economic Theory* 53: 213-38.

Teulings, C.N., I.V. Ossokina and H.L.F. de Groot (2014) Welfare benefits of agglomeration and worker heterogeneity. CESIFO Working Paper N°4939.

Thomas, I. (2002) *Transportation Networks and the Optimal Location of Human Activities: A Numerical Geography Approach.* Cheltenham, U.K.: Edward Elgar.

Topa, G. and Y. Zenou (2015) Neighborhood versus network effects. In: G. Duranton, J.V. Henderson and W. Strange (eds.) *Handbook of Regional and Urban Economics. Volume 5*. Amsterdam: Elsevier, 561-624.

Tiebout, C.M. (1956) A pure theory of local public expenditures. *Journal of Political Economy* 64: 416-24.

Ushchev, P. I. Sloev and J.-F. Thisse (2015) Do we go shopping downtown or in the 'burbs? *Journal of Urban Economics* 85: 1-15

Van Dender, K. (2003) Transport taxes with multiple trip purposes. *Scandinavian Journal of Economics* 105: 295-310.

Vickrey, W.S. (1969) Congestion theory and public investment. *Papers and Proceedings of the Eighty-first Annual Meeting of the American Economic Association*: 251-60.

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

Weber, A. (1909) *Über den Standort der Industrien.* Tübingen: J.C.B. Mohr. English translation: *The Theory of the Location of Industries.* Chicago: Chicago University Press, 1929.

Wellisch, D. (2000) *Theory of Public Finance in a Federal State*. Cambridge: Cambridge University Press.
Wildasin, D.E. (1991) Income redistribution in a common labor market. *American Economic Review* 81: 757-74.
Williamson, J.G. (1965) Regional inequality and the process of national development: A description of the patterns. *Economic and Cultural Change* 13: 1-84.
Wilson, A.G. (1970) *Entropy in Urban and Regional Modelling*. London: Pion.
Winston C., Mannering F. (2014), Implementing technology to improve public highway performance: A leapfrog technology from the private sector is going to be necessary. *Economics of Transportation* 3, 158-165
Wolff, H. (2015) Keep your clunker in the suburb: Low emission zones and adoption of green vehicles. *Economic Journal*, forthcoming.
Zeng, D.-Z. (2014) The role of country size in spatial economics: A survey of the home market effects. *Proceedings of Rijeka Faculty of Economics: Journal of Economics and Business* 32: 379-403.
Zeng, D.-Z. and T. Uchikawa (2014) Ubiquitous inequality: The home market effect in a multicountry space. *Journal of Mathematical Economics* 50: 225-33.
Zenou, Y. (2009) *Urban Labor Economics*. Cambridge: Cambridge University Press.
Zipf, G.K. (1949) *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison–Wesley Press.