Toward a unified theory of economic geography and urban economics*

Jacques-François Thisse†

26 April 2009

Abstract

In this paper, I pursue two objectives. First, I propose a primer in economic geography relying on a simple model that can be solved analytically by undergraduate students. Second, I briefly discuss two topics that, in my opinion, should rank high on the research agenda.

1 Introduction

Location theory is one of the pillars of regional science. It can be divided into three subfields: (i) spatial competition theory (Hotelling, 1929), (ii) urban economics (Alonso, 1964) and (iii) economic geography (Krugman, 1991). These domains can be distinguished through the following main feature. In spatial competition theory, consumers are immobile and travel to firms, which choose their locations strategically. In urban economics, workers compete for land and travel to the city's central business district where firms are exogenously located. In economic geography, firms and consumers are mobile while commodities are shipped across regions or countries.

These various strands of literature are often thought of as being designed to cope with location decisions made at different spatial scales. Needless to say, urban economics focuses on cities, while economic geography deals with interregional (or international) issues. This distinction is somewhat arbitrary because a national economy is often dominated by its urban system, thus making it hard to distinguish them. To make things fuzzier, spatial competition theory can address a fairly large spectrum of issues: a continuum of locations may represent everything from consumers' locations along Main Street to a large number of regions supplied by a few big firms acting strategically. In the latter context, "shopping" is replaced by "shipping," thus enabling this setting to study firms delivering their products to distant and dispersed customers. Fujita and Thisse (2002) have shed light on these various bodies of knowledge by studying the commonalities of the economics of agglomeration from the local to the global. They show that the main forces at work in these three domains are, to a large extent, fairly similar.

There are differences, however. For example, the distinctive feature of a city is its very high density of activity, which allows agents to be close to one another. Households and firms seek

*I thank Kristian Behrens, Gilles Duranton, Carl Gaigné, Gianmarco Ottaviano, and Joe Tharakan for their comments and suggestions. All of them have been my students and the time I shared with them on this occasion has been one of the most enjoyable of my academic life.

†CORE-Université catholique de Louvain, International School of Economics at Tbilissi (Georgia), and CEPR.

spatial proximity because they need to interact on a daily basis for a variety of economic and social reasons. 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 locations. At the interregional level, what matters is how capital and labor mobility affects the working of the product and labor markets in the regions of origin and destination. This feeds back in the earnings of factor-owners and changes the incentives to move through the level of demand in both regions.

The main issue then is to figure out how and when the agglomeration - or dispersion - of activities may arise as the unintended consequence of a myriad of decisions made by firms and workers pursuing their own interest. Spatial proximity to big markets and demanding customers still matters, although the distance between places seems gigantic to fellows of my generation. This is due to the amazingly low value that transport costs take on nowadays. This, however, does not imply the "death of distance." Even though transport costs must be positive for space to matter, one should not infer from this observation that location matters less when transport costs decrease. Quite the opposite: by making firms more footloose, lower transport costs make them more sensitive to minor differences between regions. As a result, a tiny difference may have a big impact on the spatial distribution of economic activity.

That said, there is a need for a more integrated theory of the location of activity. In order to achieve this goal, we must figure out *how the main forces acting at each spatial scale interact to generate the space-economy.* The outcome often looks different from the urban and interregional viewpoints, especially when we come to the design of regional and urban policies. Because such an objective is too ambitious to be reached within the format of this paper, I will confine my discussion to the following points. First, I will show that some of the key-results of economic geography may be derived within a simple framework that can be solved analytically by using math taught to undergraduate students. Section 2, therefore, will be a "primer" of sort in economic geography. In Section 3, I will start from the economic geography standpoint to provide a birds-eye view of two of the main issues that should be addressed to make this body of research closer to and more consistent with urban economics. I will refrain from discussing the relationships between economic geography and spatial competition since this has been done, among others, by Fujita and Thisse (2002).

2 A primer in economic geography

The canonical model of economic geography involves two sectors (the M-sector and the I-sector), two regions (East and West), and two production factors (the M-factor and the I-factor). Economic geography thus shares with trade theory the $2 \times 2 \times 2$ framework. There is a mass $S = M + I$ of individuals. The M-sector is formed by m firms producing an homogenous good under imperfect competition and increasing returns: each firm needs one unit of the M-factor to enter the market and produce the M-good.[1] The M-factor is *mobile* between regions; its market is perfect and global. Each of the M individuals is endowed with one unit of the M-factor; the endogenous share $\lambda \geq 1/2$ of this factor is located in West. The I-sector produces another homogeneous good under perfect competition and constant returns to scale. The I-factor is *immobile*; its markets are perfect but local. Each of the M individuals is endowed with one unit of the M-factor; the exogenous share $\theta \geq 1/2$ of this factor is located in West.[2]

[1] Formally, n is an integer. For simplicity, however, I will treat n as a real number.

[2] Throughout the paper, I will focus on West but it should be kept in mind that mirror images also holds for East.

Preferences are identical across consumers and given by a quasi-linear utility embodying a quadratic sub-utility:

$$U = \left(1 - \frac{q}{2}\right) q + q_0 \tag{1}$$

where q is the consumption of the M-good and q_0 the consumption of the numéraire.

Our purpose being to investigate how firms belonging to the M-sector distribute themselves between East and West, this effect is isolated by working with a setting in which the price of the I-good is the same across regions. This can be achieved by assuming that trading the I-good is costless. This good is chosen as the numéraire so that the price of the I-factor is 1. In contrast, the output of the M-sector is traded at a cost of τ units of the numéraire per unit shipped between the two regions.

Economic geography is dominated by two models: (i) the *footloose capital* model in which capital is the M-factor and the *mobile labor* model in which labor is the M-factor. In what follows, I first describe the market equilibrium for a given distribution of firms.

2.1 The pattern of trade

Utility maximization leads to the individual inverse demand $p = 1 - q$ for the M-good. Denoting by $s_{_W}$the mass of consumers in West, the aggregate inverse demand is given by

$$p_{_W} = 1 - \frac{Q_{_W}}{s_{_W}}$$

where $Q_{_W}$ is the total quantity of the M-good sold in this region.

Because they are spatially separated, the two regional markets are supposed to be segmented. This means that each firm chooses a specific quantity to be sold on each market; Let $q_{_{WR}}$ be the quantity of the M-good that a W-firm sells in region $R = E, W$. Market clearing for the M-good in West implies $Q_{_W} = \lambda m q_{_{WW}} + (1 - \lambda) m q_{_{EW}}$. The operating profits earned by a W-firm are as follows:

$$\pi_{_W} = q_{_{WW}} p_{_W} + q_{_{WE}} \left(p_{_E} - \tau\right).$$

The market outcome is given by a Nash equilibrium of the game in which the m firms compete in quantity. Solving the first-order conditions for profit maximization under two-way trade yields the equilibrium quantities sold by a W-firm:

$$q^*_{_{WW}} = s_{_W} p^*_{_W} \qquad \text{and} \qquad q^*_{_{WE}} = s_{_E} \left(p^*_{_E} - \tau\right)$$

where the equilibrium price in West is equal to

$$p^*_{_W} = \frac{1}{m+1} + \tau \frac{(1-\lambda)m}{m+1}. \tag{2}$$

Thus, the market price in West decreases with the number of firms located therein because this market is more competitive. In addition, this price falls when trade costs decrease because inter-regional competition gets fiercer. These results agree with what we know in spatial competition theory. Last, since consumers pay the same price regardless of the suppliers' location, the price at which the good is sold in West does not reflect the cost of supplying it, meaning there is full absorption of trade costs by exporting firms. As the same holds for the other region, we have both

intra-industry trade and *reciprocal dumping* (Brander and Krugman, 1983). Because W-firms have a lower marginal delivery cost than E-firms, they have a larger market share in West, and vice versa in East.

It remains to check when two-way trade occurs. It is readily verified that p_{W}^{*} exceeds τ for all values of $\lambda \geq 1/2$ if and only if

$$\tau < \tau_{trade} \equiv \frac{1}{m+1}. \tag{3}$$

Hence, firms choose to export when m (or τ) is small because competition on the external market is soft (or shipping the good to the external market is cheap).

Finally, using the properties of linear demand functions, it is readily verified that the equilibrium operating profits of a firm established in West may be written as follows:

$$\pi_{W}^{*} = s_{W}\left(p_{W}^{*}\right)^{2} + s_{E}\left(p_{E}^{*} - \tau\right)^{2}.$$

Thus, lowering trade costs has a negative impact on local profits but a positive impact on external profits.

2.2 The footloose capital model

Capital is the M-factor, the mass of which is K. Since one unit of capital is needed for a firm to operate, capital market clearing implies that $m = K$. To rule out comparative advantage à la Heckscher-Ohlin, each region is assumed to have the same share of capital and I-factor ($s_{W} = \theta S$). Hence, the equilibrium profits made by a firm located in West are equal to

$$\pi_{W}^{*} - r_{W} = \theta S\left(p_{W}^{*}\right)^{2} + (1-\theta)S\left(p_{E}^{*} - \tau\right)^{2} - r_{W}$$

where r_{W} is the rental rate of capital in West. Profits are distributed to consumers and are supposed to be sufficiently large for the consumption of the numéraire to be positive at the equilibrium outcome.

Capital-owners seek the highest rental rate while firms aim to attract capital through a competitive bidding process that ends when no firm can make positive profits at the equilibrium market prices. This means that all operating profits go to the capital-owners. Hence, a spatial equilibrium is reached at $1/2 \leq \lambda^{*} < 1$ when the equilibrium operating profits are the same in the two regions and equal to the common rental rate:

$$r_{W}(\lambda) = \pi_{W}^{*}(\lambda) = \pi_{E}^{*}(\lambda) = r_{E}(\lambda).$$

Otherwise capital is fully agglomerated in West ($\lambda^{*} = 1$) since $r_{W}(1)$ exceeds $r_{E}(1)$.

Substituting the equilibrium prices (2) into $\pi_{W}^{*}(\lambda)$ and $\pi_{E}^{*}(\lambda)$ yields the following profit differential:

$$\pi_{W}^{*}(\lambda) - \pi_{E}^{*}(\lambda) \propto \left[2\left(1 - \frac{\tau}{2}\right)\theta - \left(1 - \frac{m+1}{2}\tau\right) - \lambda m\tau\right].$$

The equation $\pi_{W}^{*}(\lambda) - \pi_{E}^{*}(\lambda) = 0$ has a unique solution λ^{*}, which is the equilibrium distribution of firms:

$$\lambda^{*} - \frac{1}{2} = \frac{2-\tau}{m\tau}\left(\theta - \frac{1}{2}\right) > \theta - \frac{1}{2}.$$

Using (3), λ^{*} exceeds θ as long as $\theta > 1/2$ since $\lambda^{*}(1/2) = 1/2$ and $\partial\lambda^{*}/\partial\theta > 1$. This shows the presence of a *home market effect*: the larger region hosts a more than proportionate share of

firms (Krugman, 1980). Furthermore, we have $\partial\lambda^*/\partial\tau < 0$. To put it differently, lower trade costs leave the agglomeration force unchanged but weakens the dispersion force, thus leading to a more concentrated pattern of firms in West. Finally, all firms set up in West when the two regions have very different sizes because $\lambda^*(1) > 1$.

In a nutshell, size matters for the spatial distribution of capital once competition is imperfect and trade costs are positive: *regions and countries tend to specialize and export goods for which they have a large local market.*

2.3 The mobile labor model

Labor is the M-factor, the mass of which is L. Since one unit of labor is needed to operate one firm, labor market clearing implies $m = L$. In order to control for any comparative advantage, the owners of the I-factor are now equally split between the two regions ($s_{W} = I/2 + \lambda L$). Hence, the equilibrium profits of a firm located in West are defined as follows:

$$\pi^*_{W} - w_{W} = (I/2 + \lambda L)\left(p^*_{W}\right)^2 + [I/2 + (1-\lambda)L]\left(p^*_{E} - \tau\right)^2 - w_{W}$$

where w_{W} is the wage rate in West. Observe the difference between the footloose capital and mobile labor models: market sizes are fixed in the former (θS), but variable in the latter ($I/2 + \lambda L$).

Workers choose the region that maximizes their indirect utility evaluated at the equilibrium prices and wages:[3]

$$V = CS + w.$$

The individual consumer surplus in West is given by

$$CS^*_{W} = \frac{m^2}{2\,(m+1)^2}\left[1 - \tau(1-\lambda)\right]^2$$

which increases with the share of firms located in West and decreases with the level of trade costs. Regarding the equilibrium wage rate prevailing in West, it is obtained through a bidding process that ends when no firm can earn strictly positive profits at the equilibrium market prices. In other words, all operating profits go to the workers:

$$w_{W} = \pi^*_{W}.$$

Because workers seek the highest utility level, their migration is governed by the utility differential: they flow from East to West when $V_{W} - V_{E} > 0$ and stay put when $V_{W} - V_{E} = 0$. Thus, a spatial equilibrium is reached at $1/2 \leq \lambda^* < 1$ when $V_{E}(\lambda^*) = V_{W}(\lambda^*)$; this equilibrium is stable when a marginal change in λ leads workers back to the initial configuration. An agglomerated equilibrium ($\lambda^* = 1$) being a corner point, it is stable whenever it exists.

One of the main differences regarding the mobility of labor and capital is that the former is driven by both their nominal wage and market price whereas the latter is governed by its nominal rate of return only. Increasing the number of firms located in West makes local consumers better off because the local market price is lower. Regarding the evolution of wages, such an increase in the number of firms has two opposite effects. First, as more firms locate in West, we have just seen that the local price decreases. This yields lower operating profits, but this effect is weaker when τ

[3] For simplicity, I assume that wages are sufficiently high for workers to consume the numériare.

is lower because $\partial^2 p_W^*(m,\tau)/\partial m \partial \tau < 0$. However, this negative competition effect is not the only one at work. For some firms to move to West, some workers have to follow suite. What is at work here is a positive *market expansion effect* generated by the linkage between firms' locations and workers' expenditures, which leads to higher operating profits. Since the two effects oppose each other, the net impact on wages is therefore a priori ambiguous, although low trade costs seem to favor agglomeration.

To confirm this idea, we need a more formal argument. To this end, we plug the equilibrium prices and wages into $V_W - V_E$, which then depends upon λ only. After some simple manipulations, we obtain the following expression:

$$V_W(\lambda) - V_E(\lambda) = CS_W - CS_E + w_W - w_E \propto \tau(\tau^* - \tau) \cdot (\lambda - 1/2) \tag{4}$$

where

$$\tau^* = \frac{2(2+3L)}{2+2I+5L+2IL+2L^2}.$$

It follows immediately from (4) that the symmetric configuration ($\lambda = 1/2$) is always a spatial equilibrium. The utility differential has the same sign as $(\lambda - 1/2)$ if and only if τ is smaller than τ^*; otherwise it has the opposite sign. Hence, for large trade costs, that is, when τ exceeds τ^*, $\lambda^* = 1/2$ is the only stable equilibrium. In stark contrast, when $\tau < \tau^*$, the symmetric equilibrium becomes unstable because $V_W(\lambda) - V_E(\lambda) > 0$ for all $\lambda > 1/2$. In this case, all firms and workers agglomerate in West ($\lambda^* = 1$), which is now the core of the economy, while East is the periphery. To sum up: *high trade costs yields dispersion, whereas low trade costs leads to agglomeration*, as in Krugman (1991) and Fujita *et al.* (1999).

Another important implication of the model is that it generates a *putty-clay geography*. If firms and workers can settle in East just as well as in West, then once the agglomeration process is set into motion it keeps developing in the same region. Individual choices become more rigid because of the self-reinforcing nature of the process of agglomeration, which sparks a lock-in effect. Therefore, the magnet region is a priori indeterminate in that its selection is likely to depend on small events. In the past, the resource endowment was often the main determinant for the location of firms. To a large extent, this explains the difference observed between the old and new geographies of production. One should also add that the above model provides a rational for the (relative) inertia displayed by urban systems.

Note that $\tau^* < \tau_{trade}$ holds provided that the mass I of immobile consumers is large while the mass L of mobile consumers is low. Otherwise, the demand from East will always be too low to prevent the full agglomeration of the M-sector in West. For example, assuming $I = 1$ implies that τ^* exceeds τ_{trade} for all $L > 0$. Furthermore, when the mass L of workers - or, equivalently, the number m of firms - is arbitrarily large, both τ^* and τ_{trade} converge to 0, which means that there is no trade and that the M-sector is always dispersed. This is so because a very large number of firms amounts to assuming that the M-sector displays constant returns, which implies that each region becomes an autarky. By contrast, when L is small, few firms operate because scale economies are strong. In this case, firms and workers co-locate in West provided that trade costs are sufficiently low, thus confirming the importance of increasing returns for the formation of urban agglomerations.

2.4 A short synthesis: part I

In many countries, there are sizable and persistent spatial variations in population sizes, average incomes, regional production structures, the cost of living, and the distribution of jobs. All these

magnitudes are endogenous and the values they take are not imposed by nature. On the contrary, they are determined by the interaction between markets, public policies, and the mobility of production factors. It is the spatial facet of these numerous interactions that forms the realm of economic geography.

That said, the message of this primer is clear: *market integration and factor mobility exacerbate regional disparities*. This result has far-reaching implications. Once we account for the mobility of some factors, market integration or transport-improving policies could well yield more spatially imbalanced patterns of activities, thus providing a rationale for the above-mentioned disparities. Economic geography thus delivers a message very different from standard trade theory, which predicts regional convergence with respect to factor prices and earnings. The above two models are not identical, however, and they do not deliver exactly the same message. The main difference is that the process of agglomeration is smooth in the footloose capital model, while it displays a bang-bang behavior in the mobile labor model. In other words, *capital mobility and labor mobility are not equivalent for the space-economy*. Furthermore, agglomeration stems from exogenous market size differences in the footloose capital model. By contrast, in the mobile labor model, agglomeration is the outcome of interactions between the product and labor markets in an otherwise symmetric environment.

My last point is methodological. The model proposed here is very simple to manipulate and could, therefore, be used as a building-block in frameworks addressing broader issues.

3 What next?

The main models developed in economic geography overlook many costs whose origin lies in the space-economy (for example, the various congestion costs generated by the emergence of an agglomeration) and, conversely, overlook other agglomeration economies, such as a better matching on labor markets, the proximity of intermediate inputs and the existence of local knowledge spillovers. Along the same line, it is hard to see why trading the I-good is costless in a model seeking to ascertain the overall impact of trade costs on the location of economic activities. In addition, these models only account for two sectors and two regions. Even this is not really so since the I-sector is given a very restricted role, its job being to guarantee the equilibrium of the trade balance. Although these issues have already attracted some attention, it seems fair to say that most of these contributions lack a common thread that would help the development of new and original research.

3.1 On the interactions between the local and the global

To begin, I want to discuss some relationships between economic geography and urban economics. When economic geography came onto the stage, urban economics was already a well-established field, starting with what was called the "New Urban Economics" in the 1970s. Since then, we have learned a lot and it seems fair to say that urban economics is much ahead of economic geography. Yet, economic geographers might have something important to add to the tool-box of urban economists, that is, *the interaction between the local and the global*. This is because market integration is likely to affect the attractiveness of cities. Specifically, it seems natural to study how their competitiveness within an interregional economic space depends on their size and internal structure, which affect trade flows and, conversely, how trade influences the organization of cities and the interregional location of firms. The structure of cities also has a strong impact on the working of local labor and housing markets, which therefore affects workers' decisions to migrate.

All of this can be achieved in a model that encapsulates both trade costs and urban costs, that is, land rent augmented by commuting costs. That said, the development of modern information and communication technologies is a new major force that facilitates job decentralization. Therefore, in addition to trade and commuting costs, we must also account for communication costs.

In what follows, I graft a standard urban economics model onto the mobile labor model. To achieve my goal, I assume that there are no immobile consumers ($I = 0$), thus suppressing the dispersion force considered by Krugman. On the other hand, I assume that the concentration of workers in West generate costs that rise with the size of the local population, a new dispersion force suggested by Tabuchi (1998). From now on, each region is described by a one-dimensional bounded space, which can accommodate both firms and workers. Whenever a city exists, it has a central business district (CBD) located at $x = 0$ where an endogenous number of firms choose to set up for reasons surveyed by Duranton and Puga (2004). Workers consume a unit residential plot and share the same utility (1). Each worker bears a unit commuting cost $t > 0$. Consequently, a worker located at a distance x from the W-CBD has a budget constraint given by

$$q_{W} p_{W} + q_0 + R_{W}(x) + tx = w_{W}$$

and pays a land rent equal to

$$R_{W}(x) = t\left(\frac{\lambda L}{2} - x\right).$$

Her indirect utility is now given by

$$V_{W}(\lambda) = CS_{W} + w_{W} - \frac{t}{2}\lambda L$$

which implies

$$V_{W}(\lambda) - V_{E}(\lambda) \propto [(a - b\tau)\tau - t]\left(\lambda - \frac{1}{2}\right) \tag{5}$$

where $a \equiv 2(3m+2)/(m+1)^2$ and $b \equiv (m+2)(2m+1)/(m+1)^2$.[4] Note that $a - b\tau > 0$ because $\tau_{trade} < b/a$.

It follows from (5) that the stability of the symmetric configuration depends on the relative values of trade and commuting costs. More precisely, inspecting $(a - b\tau)\tau - t$ reveals that the M-sector is agglomerated into a single monocentric city when commuting costs are low, trade costs are high, or both. This is very different from what we have found in the foregoing section since high trade costs now trigger the agglomeration of the M-sector, as in Helpman (1998). This difference in results may be explained as follows. The intensity of the dispersion force generated by the demand of immobile consumers is unaffected by the agglomeration of the M-sector. In contrast, when this sector gets more and more concentrated in West, the intensity of the dispersion force generated by the existence of urban costs keeps growing. When trade costs are low, this force overcomes the agglomeration force, thus leading to the dispersion of the M-sector between different cities. However, lowering commuting costs can sustain the agglomeration as an equilibrium by reducing urban costs. In other words, the parameter t provides a simple measure of the intensity of the dispersion force. All of this suffices to show how some global forces may affect the structure of the urban system and the size of cities, and vice versa.

The story does not end here, however. The foregoing argument shows that workers and firms get dispersed because urban costs are high in West. Once it is recognized that big cities may become polycentric through the development of secondary business centers (SBDs), the average

[4] For simplicity, I assume that the aggregate land rent is distributed to absentee landlords.

commuting costs and land rent borne by those working in a SBD are lower than those paid by the individuals working in the CBD. Simultaneously, because fewer workers commute to the CBD, the corresponding workers also bear lower urban costs. In sum, workers' welfare becomes higher in West when the corresponding city becomes polycentric. By the same token, firms are able to pay lower wages and land rents while retaining most of the benefits generated by large urban agglomerations. Thus, we may expect the escalation of urban costs in large cities to prompt the redeployment of activities in a polycentric pattern. For this to happen, however, firms located in SBDs must be able to maintain very good access to the inner city, which provides highly specialized business-to-business services (Porter, 1999), which in turn requires low communication costs.

By introducing communication costs, we account for the fact that agglomeration and dispersion across space may take two quite separate forms because they are now compounded by centralization or decentralization of activities within the same city. When commuting and communication costs are high, the space-economy is likely to be formed by several small cities. In contrast, when communication costs reach low values while commuting costs take intermediate values, large polycentric cities would emerge. Therefore, by facilitating the formation of SBDs, the development of new information and communication technologies slows down the redispersion process. Stated differently, *decentralization within the metropolis allows the core regions to retain their primacy* (Cavailhès *et al.*, 2007). Such results shed light on the interplay between different types of spatial friction affecting the location of economic activities *between* and *within* urban agglomerations. It also draws attention to two facts that policy makers often neglect: on the one hand, local factors may change the global organization of the economy and, on the other, global forces may affect the local organization of production and employment. Among other things, this relationship calls for a better coordination of transport policies at the city and interregional level.

Ever since Alfred Marshall, it is well known that the clustering of activities allows firms to become more productive thanks to the various types of externalities linked to the size of the workforce. I want to single out here the work of Abdel-Rahman and Fujita (1990) who provide a very compelling argument for the existence of such externalities. They show how a larger population of workers yields a finer division of labor in the intermediate sector, which translates into an increasing marginal productivity of labor in the final sector. Their model can be combined with the foregoing one to show how the redispersion process is braked because it would lead to a decrease in firms' productivity.

It should also be recognized that globalization has two facets in that it goes hand in hand with lower trade costs between places as well as with lower communication costs within firms, which reduces the costs of coordinating production across space and time zones. Otherwise, how do we explain why the relocation of manufacturing activities did not occur earlier in Asia where wages were very low for a long time? More precisely, low communication and trade costs permit firms to organize and perform discrete activities in distinct locations, which altogether form the supply chain starting at the conception of the product and ending at its delivery. This spatial fragmentation of production aims at taking advantage of differences in technologies, factor endowments, or factor prices across places. Combined with low communication costs, trade liberalization has an impact that vastly differs from what we have seen in Section 2: firms no longer perform all their activities under the same roof. Assuming that firms have both headquarters and plants, decreasing trade *and* communication costs lead to a pattern of activity in which headquarters remain in the core together with high-level business-to-business services, while plants move to the periphery (Fujita and Thisse, 2006). In contrast, when communication costs are high, plants and headquarters locate side-by-side, while low trade costs fosters the agglomeration of vertically integrated firms.

Compiling these ideas, we may conclude that *the decentralization of jobs may occur both in*

the small and in the large. Where jobs are created and destroyed depend on the components of the supply chain, some having to be within close proximity, whereas others can be located at a distance. This points to the need for urban economists and regional scientists to pay more attention to modern theories of the firm in order to build relevant models of the urban system (Spulber, 2009).

3.2 How many?

Let me now turn to the so-called "dimensionality problem". This is not a brand new issue for trade people. For example, Deardorff (1984, p. 468) wrote several years ago that "...the Heckscher-Ohlin theorem is derived from a model of only two of each of goods, countries, and factors of production. It is unclear what the theorem says should be true in the real world where there are many of all three." Under some proviso, this theorem has been extended to the case of many goods and factors. The case of many countries remains largely unexplored, however, especially in the presence of trade costs. The same holds true for economic geography. On the other hand, Henderson (1974, 1988) has developed a compelling and original approach that allows him to describe how an urban system involving an endogenous number of cities of different sizes and industrial compositions may emerge. For our purpose, the issue is that Henderson's cities are like floating islands because shipping commodities across cities is assumed to be costless. Or, to put it differently, we do not know where cities are located, hence what their relative positions are. These issues are hard to study under the assumption of zero trade costs, unless we consider a space endowed with heterogeneous natural endowments. Glaeser has forcefully argued that inter-city trade costs are negligible and may, therefore, may be disregarded. This might well be so within the American space-economy. However, one should keep in mind that trade costs do not subsume to the sole monetary cost of shipping goods (Anderson and van Wincoop, 2004). In addition, what is true for the US need not be true for the rest of the world, as shown by all the estimations of the gravity equation (Disdier and Head, 2008).

But why should one bother about the existence of many regions instead of two? The new fundamental ingredient that a multi-regional setting brings about is that *the accessibility to spatially dispersed markets varies across regions*. In other words, the relative position of a region within the network of exchanges (which involves cultural, linguistic and political proximity) matters. Any global change in this network such as market integration is likely to trigger complex effects that vary in non-trivial ways with the properties of the graph representing the network (Behrens *et al.*, 2007). When there are only two regions, the overall impact can be captured through the sole variation in the cost of trading goods between them. On the contrary, when there are more than two regions, a change that directly affects two regions generates general equilibrium effects that are unlikely to leave the remaining regions unaffected. In particular, a multi-regional setting should make it possible to study how lowering trade costs amplify or reduce the geographical advantage and disadvantage held by different regions (Matsuyama, 1998).[5]

Economic geography and urban economics do not have much to say regarding those questions, although the evidence shows that accessibility strongly affects the potential of regions and cities for development (Collier, 2007). To illustrate, Limão and Venables (2001) show that, in comparison with the median coastal country, the median landlocked country bears an additional transport cost of 55%, while its volume of trade at the same income level and distance decreases by 60%. With such figures in mind, it should be clear that accounting explicitly for a multi-regional economy with

[5] To be sure, there are papers dealing with several regions or countries, but they typically assume that locations are equidistant. This can hardly be considered as a true extension of the two-region setting. A noticeable exception is Eaton et Kortum (2002) who provide a new and general framework that should allow one to derive further results.

different trade costs should rank high on the research agenda.

Working with many sectors is very important too, because industries differ in terms of scale economies, product differentiation and trade costs. It is the interaction between all these parameters that determines how labor and capital are distributed across places and sectors. To be precise, the activities located in one region must be studied in relation to those undertaken in others. Unfortunately, the task does not seem to be easy. The reason lies in the creation and destruction of equilibria (the multiplicity of equilibria) that makes it hard to predict which path the economy will follow during the integration process, even in very simple settings (Tabuchi and Thisse, 2006). Such issues are endemic to problems involving scales economies. In order to build sensible selection rules, we must graft learning or evolutionary processes onto economic geography models. The assumption of perfect foresights can at best serve as a benchmark. Another way out way out is to assume that agents are sufficiently heterogeneous (Herrendorf *et al.*, 2000).

To sum up, the dimensionality problem appears to be even tougher in economic geography than in trade theory. This need not be a hopeless goal, however. It is well known that physics provides solutions to the two-body problem and to continuum mechanics, which has a very large number of infinitesimal bodies, but not for numbers of bodies falling in between. Therefore, a possible way out to the dimensionality problem is the use of infinitely many sectors and regions/cities. In this case, each one of them is negligible to the global economy. Getting rid of interactions that remain out of our reach could yield spatial and economic structures amenable to analytical solutions.[6] To illustrate, monopolistic competition, which involves a continuum of firms when it is properly modeled, may be reinterpreted as a setting involving a continuum of sectors, each being endowed with a single firm.

3.3 A short synthesis: part II

The above discussion was not intended to be comprehensive. It deals with what I consider as some of the main issues that should be addressed in the near future. First, we need a better integration of different types of spatial frictions in order to figure out how forces acting on different spatial scales shape the global economic landscape. This task must be accomplished within a multi-city framework in order to capture most general equilibrium effects. Three is already better than two. In doing so, however, we cannot ignore the relative position of cities, as expressed by different accessibility measures such a trade or transport costs. If we academics wonder about *why* cities exist and grow (or shrink) and why sizable and persistent gaps between and within countries occur, what people and policy-makers care about is *where* these things happen.

[6] Note that Neary (2003) has proposed to work with a continuum of oligopolistic sectors in a two-country setting with zero trade costs.

References

[1] Abdel-Rahman, H. and M. Fujita. 1990. "Product Variety, Marshallian Externalities, and City Sizes," *Journal of Regional Science*, **30**, 165-83.

[2] Alonso, W. 1964. *Location and Land Use*. Cambridge, MA: Harvard University Press.

[3] Anderson, J. and E. van Wincoop. 2004. "Trade Costs," *Journal of Economic Literature*, **XLII**, 691-751.

[4] Behrens, K., A.R. Lamorgese, G.I.P. Ottaviano and T. Tabuchi. 2007. "Changes in Transport and Non-transport Costs: Local vs Global Impacts in a Spatial Network," *Regional Science and Urban Economics*, **37**, 625-648.

[5] Brander, J. and P.R. Krugman. 1983. "A 'Reciprocal Dumping' Model of International Trade," *Journal of International Economics*, **15**, 313-321.

[6] Cavailhès, J., C. Gaigné, T. Tabuchi, and J.-F. Thisse. 2007. "Trade and the Structure of Cities," *Journal of Urban Economics*, **62**, 383-404.

[7] Collier, P. 2007. *The Bottom Billion. Why the Poorest Countries Are Failing and What Can Be Done About It*. Oxford: Oxford University Press.

[8] Deardorff, A.V. 1984. "Testing Trade Theories and Predicting Trade Flows," in R.W. Jones and P.B. Kenen (eds), *Handbook of International Economics. Volume 1*. Amsterdam: North Holland, pp. 467-517.

[9] Disdier, A.C. et K. Head. 2008. "The Puzzling Persistence of the Distance Effect on Bilateral Trade," *Review of Economics and Statistics*, **90**, 37–48.

[10] 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. Volume 4*. Amsterdam, North Holland, pp. 2063-2117.

[11] Eaton, J., and S. Kortum. 2002. "Technology, Geography and Trade," *Econometrica*, **70**, 1741-1780.

[12] Fujita, M., P. Krugman and A.J. Venables. 1999. *The Spatial Economy. Cities, Regions and International Trade*. Cambridge, MA: The MIT Press.

[13] Fujita, M. and J.-F. Thisse. 2002. *Economics of Agglomeration. Cities, Industrial Location and Regional Growth*. Cambridge: Cambridge University Press.

[14] Fujita, M. and J.-F. Thisse. 2006. "Globalization and the Evolution of the Supply Chain: Who Gains and Who Loses?," *International Economic Review*, **47**, 811-836.

[15] 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, pp. 33-54.

[16] Henderson, J.V. 1974. "The Sizes and Types of Cities," *American Economic Review*, **64**, 640-656.

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

[18] Hotelling, H. 1929. "Stability in Competition," *Economic Journal*, **39**, 41-57.

[19] Herrendorf, B., A. Valentinyi, and R. Waldmann. 2000. "Ruling out Multiplicity and Indeterminacy: The Role of Heterogeneity," *Review of Economic Studies*, **67**, 295-307.

[20] Krugman, P.R. 1980. "Scale Economies, Product Differentiation, and the Pattern of Trade," *American Economic Review*, **70**, 950-959.

[21] Krugman, P. 1991. "Increasing Returns and Economic Geography," *Journal of Political Economy*, **99**, 483-499.

[22] Limao, N. and A. Venables. 2001. 'Infrastructure, Geographical Disadvantage, Transport Costs, and Trade," *World Bank Economic Review*, **15**, 451-479.

[23] Matsuyama, K. 1998. ,Geography of the World Economy," Mimeo, Northwestern University, http://www.kellogg.nwu.edu/research/math

[24] Neary, P. 2003. "Globalization and Market Structure," *Journal of the European Economic Association*, **1**, 245-271.

[25] Porter, M.E. 1995. "Competitive Advantage of the Inner City," *Harvard Business Review*, May-June, 55–71.

[26] Spulber, D. 2009. *The Theory of the Firm: Microeconomics with Endogenous Entrepreneurs, Firms, Markets, and Organizations*. Cambridge: Cambridge University Press.

[27] Tabuchi, T. 1998. "Urban Agglomeration and Dispersion: A synthesis of Alonso and Krugman," *Journal of Urban Economics*, **44**, 333-51.

[28] Tabuchi, T. and J.-F. Thisse. 2006. "Regional Specialization, Urban Hierarchy, and Commuting Costs," *International Economic Review*, **47**, 1295-1317.