DISCUSSION PAPER SERIES

No. 3838

AGGLOMERATION AND ECONOMIC GEOGRAPHY

Gianmarco I P Ottaviano
and Jacques-François Thisse

INTERNATIONAL TRADE

Centre for Economic Policy Research

www.cepr.org

Available online at: **www.cepr.org/pubs/dps/DP3838.asp**

ISSN 0265-8003

AGGLOMERATION AND ECONOMIC GEOGRAPHY

Gianmarco I P Ottaviano, Università di Bologna and CEPR
Jacques-François Thisse, CERAS, Paris; CORE, Université Catholique de Louvain and CEPR

Discussion Paper No. 3838
March 2003

Centre for Economic Policy Research
90–98 Goswell Rd, London EC1V 7RR, UK
Tel: (44 20) 7878 2900, Fax: (44 20) 7878 2999
Email: cepr@cepr.org, Website: www.cepr.org

CEPR Discussion Paper No. 3838

March 2003

ABSTRACT

Agglomeration and Economic Geography*

Peaks and troughs in the spatial distributions of population, employment and wealth are a universal phenomenon in search of a general theory. Such spatial imbalances have two possible explanations. In the first, uneven economic development can be seen as the result of the uneven distribution of natural resources. This is sometimes called ‘first nature’ and refers to exogenously given characteristics of different sites. It falls short of providing a reasonable explanation of many other clusters of activities, however, which are much less dependent on natural advantage. The aim of geographical economics is precisely to understand what are the economic forces that, after controlling for first nature, account for ‘second nature’, which emerges as the outcome of human beings’ actions to improve upon the first one. Specifically, geographical economics asks what are the economic forces that can sustain a large permanent imbalance in the distributions of economic activities. In this Paper, we focus on the so-called ‘new economic geography’ approach. After having described some of the main results developed in standard location theory, we use a unified framework to survey the home market effect as well as core-periphery models. Geographers have criticized these models because they account for some spatial costs while putting others aside without saying why. Furthermore, core-periphery models also exhibit some extreme features that are reflected in their bang-bang outcomes. We thus move on by investigating what the outcomes of core-periphery models become when we account for a more complete and richer description of the spatial aspects that these models aim at describing. We conclude by suggesting new lines of research.

JEL Classification: F12, F16 and R12
Keywords: agglomeration, economic geography, increasing returns, monopolistic competition and transport costs

Gianmarco I P Ottaviano
Università di Bologna
Dip Scienze Economiche
Piazza Scaravilli 2
40126 Bologna
ITALY
Email: ottavian@economia.unibo.it

Jacques-François Thisse
CORE
Université Catholique de Louvain
34 Voie du Roman Pays
B-1348 Louvain-la-Neuve
BELGIUM
Tel: (32 10) 474 312
Fax: (32 10) 474 301
Email: thisse@core.ucl.ac.be

*This Paper has been prepared for inclusion in the forthcoming Volume IV of the *Handbook of Regional and Urban Economics*, edited by V Henderson and J-F Thisse. We thank K Behrens, J Hamilton, V Henderson, T Mayer, Y Murata, E Paluzie, P Picard, D Puga, and T Tabuchi for helpful comments and discussions. This research was supported by the Ministère de l'education, de la recherche et de la formation (Communauté française de Belgique), Convention 00/05-262. Both authors are also grateful to the RTN Program of the European Commission for financial support.

Submitted 25 February 2003

1 Introduction

Peaks and troughs in the spatial distributions of population, employment and wealth are a universal phenomenon in search of a general theory. At a high level of abstraction spatial imbalances have two possible explanations.[1] First of all, uneven economic development can be seen as the result of the uneven distribution of natural resources. This is sometimes called 'first nature' and refers to exogenously given characteristics of different sites, such as the type of climate, the presence of raw materials, the proximity to natural ways of communication, etc. First nature is clearly important to explain the location of heavy industries during the Industrial Revolution, because the proximity of raw materials was a critical factor, or why Florida keeps attracting so many (retired) people. However it falls short of providing a reasonable explanation of many other clusters of activities, which are much less dependent on natural advantage (think of the metropolitan area of Tokyo or the Silicon Valley).

The aim of geographical economics is precisely to understand what are the economic forces that, after controlling for first nature, account for 'second nature', which emerges as the outcome of human beings' actions to improve upon the first one. From a methodological point of view, geographical economics starts with considering an initial situation in which space is homogenous and production activities are equally present at all sites. Then, it asks what are the forces that can allow a *small* (possibly temporary) asymmetric shock across sites to generate a *large* permanent imbalance in the distributions of economic activities. Among the various classes of models that have been put forth to address this question, we focus on the so-called *new economic geography* (in short NEG).

NEG has been initiated by three authors, namely Fujita (1988), Krugman (1991) and Venables (1996) who all use general equilibrium models with monopolistic competition. Most models in NEG assume the existence of two sectors, the 'modern' and the 'traditional' sectors. At the time of the Industrial Revolution, the modern sector was *manufacturing*. The geographical concentration of industry generated an additional demand for manufactured goods, as shown by the history of the Manufacturing Belt in the United States or the development of the Ruhr in Germany. Today, the modern sector is the *service industry* in which firms do not only supply consumers and manufacturing firms, but also serve each other. The tendency toward agglomeration is thus strengthened by the fact that business services tend to work more and more for headquarters and research labs of manufacturing firms, which remain mostly located in large urban agglomerations. Thus, our point is that 'what the two sectors are' changes with the stage of development of the economy as well as with the epoch under consideration.

The main results obtained by NEG after one decade of research have been synthesized in a book written by the same authors (Fujita *et al.*, 1999).

[1] Unevenness arises across countries in the world, across regions and cities within a country, and across neighbors within a city. In this chapter, we focus mainly on explanations dealing with the first two levels of analysis - countries and regions.

During this period, there have been endless debates involving economists, geographers and regional scientists about what was really 'new' in the NEG (see, e.g. Isserman, 1996 and Martin, 1999). By now, it is widely recognized that many ideas had been around for a long time in the works of economic geographers and location theorists.[2] We survey these ideas in section 2 and conclude that, despite several early and outstanding contributions, it is fair to say that economic geography and location theory lay for long in the periphery of mainstream economic theory. The reason for such emargination is likely to be found in the difficulty for the competitive paradigm, which has dominated so much economic research, to explain the formation of economic agglomerations. Indeed, as shown by Starrett (1978), cities, regional specialization and trade *cannot* arise at the competitive equilibrium of an economy with a homogenous space. The reason for this somehow amazing and seldom mentioned result is the presence of space-specific nonconvexities that prevent the existence of a nontrivial competitive equilibrium.

NEG may be viewed as an attempt to overcome this theoretical deadlock. It does so by connecting trade and location theories, a research target put forward several years ago by Ohlin *et al.* (1977).[3] In this perspective, modern theories of agglomeration are very much dominated by a simple principle: *a market place is the core of the economy because this place is its main market*. We illustrate this point in section 3 by means of what is called the 'home market effect' (henceforth, HME). According to Helpman and Krugman (1985, p. 197), once transport costs are explicitly accounted for, this effect arises when imperfectly competitive industries tend to concentrate their production in their larger markets and to export to smaller ones.[4] Stated differently, the HME has the features of a 'gravitational force' that attracts imperfectly competitive sectors towards larger markets. Accordingly, it allows small permanent shocks to market size to generate large permanent disparities in the location of firms. For this reason, the HME is the basic ingredient that lies at the heart of most models of agglomeration. Its historical relevance in the industrialization of Europe is emphasized by Pollard (1981) for whom, even though there are examples of export-led developments, "it is obviously harder to build up an industrial complex without the solid foundation of a home market" (p. 249). We discuss the reasons that may explain why such an effect exists and uncover its relationships with standard location theory. In particular, we argue that the market equilibrium is the outcome of the interplay between two opposing forces, a market-crowding force and a market-access force, very much as in spatial

[2]See Fujita and Thisse (2002) for an integrated overview of new economic geography and standard location theory.

[3]Note that agglomeration may arise even when transport costs are sufficiently high for trade not to occur. In the framework described in section 4.1, the spatial equilibrium pattern is then determined by the ratio of the mobile to the immobile factor: the larger this ratio, the larger the agglomeration (Behrens, 2002). Hence, contrary to general beliefs, agglomeration is not a 'by-product' of trade; it may also emerge in an autarkic world.

[4]In this chapter, transport costs are broadly defined to include all impediments caused by distance, such as shipping costs per se, tariff and non-tariff barriers to trade, different product standards, difficulty of communication, and cultural differences.

competition à la Hotelling. Using two different models of monopolistic competition, we show that, in the case of two regions, *a more than proportionate fraction of the modern sector is established within the larger region.* Very much as in standard location theory, physical capital is a disembodied factor, which can be used in either region and moves from one region to the other seeking the highest *nominal* rate of return.

The main limit of the HME is that, on its own, it is not able to explain why even small *temporary* shocks can have large permanent effects on the economic landscape. This is achieved by the models discussed in section 4. They may be subdivided into two categories according to whether or not labor is mobile. The first category assumes that part of the labor force is mobile (section 4.1). Unlike the models of section 3 in which the capital-owners repatriate all their earnings where they live, the core-periphery (in short CP) model assumes that mobile workers spend their income in the region in which they are active. In other words, the production factor is now embodied in workers and this difference suffices to give rise to circular causation in locational decisions. When transport costs are sufficiently low, such an ex-post immobility of income exacerbates the transient HME caused by a temporary shock to market size. The result is that *all firms belonging to the modern sector end up locating within a single region* (the 'core'), the other region being specialized in the traditional sector (the 'periphery'). In other words, workers' mobility strengthens much more than proportionally the initial advantage given by a larger initial market size, thus leading to the 'amplification' of the HME. By contrast, when transport costs are sufficiently high, the HME vanishes in that each region ends up with the same size. Under these circumstances, each factor gets the same earning regardless of its location, very much as in the neoclassical Heckscher-Ohlin setting. Although the idea is not new, the CP model is the first general equilibrium model that produces an uneven economic space from an otherwise even physical space as a consequence of low transport costs and mobile production factors. Compared to the HME models, one distinguishing feature of CP models is that human capital moves seeking the highest *real* rate of return.

Armchair evidence shows, however, that agglomeration arises even in the absence of labor mobility. This leads us to investigate the second category of models in which 'vertical linkages' of the type stressed by Venables (1996) is the main reason for the existence of agglomeration (section 4.2). Even though all workers stay put, the size of a regional market remains variable because of input-output linkages. The reason is that, whenever a firm sets up in a particular region, the local size of its upstream and downstream markets expand.

Models in NEG have been criticized by geographers and location theorists because they account for 'some' spatial costs while putting 'others' aside without saying why. First, transport costs of the output of the traditional sector are not zero in the real world. Second, the agglomeration of the modern sector within a region gives rise to urban costs that typically increase with the size of the corresponding population. To be sure, all these costs have declined since the beginning of the Industrial Revolu-

tion but they did not do so at the same speed. Thus, what matters for the space-economy is their relative variation over time. Furthermore, CP models share with Bertrand's setting some unpleasant and extreme features that are reflected in their 'bang-bang' outcomes: only full agglomeration and full dispersion may occur, whereas changes in the spatial pattern are always catastrophic. This does not strike us as being plausible. As argued by Hotelling (1929) long ago, such results arise because the modeling strategy is often too simple: more meaningful models should account explicitly for some heterogeneity across agents. The purpose of section 5 is to investigate what the CP model becomes when we account for a more complete and richer description of the spatial aspects that this model aims at describing.

In section 6, we suggest new lines of research. Before proceeding, a last comment is in order. The contributions in NEG have flourished during the last decade. Instead of providing a superficial overview of the field, we have chosen to concentrate on the main ideas and results within a unified framework that slightly departs from the most standard models. This framework is summarized in Table 1, whereas extensions are mentioned in footnotes where the reader may find additional references for future reading.[5]

2 The legacy of location theory

The objective of location theory is to answer the question: why some particular production activities (such as plants, offices, public facilities, etc.) choose to establish themselves in some particular places of a given space? This question can be broken in two sub-questions as to the location of a firm and the location of an industry. As to the former, one may ask: why does an isolated firm faces a location problem and how does it solve it? As to the latter, one may ask: how are firms' locational decisions intertwined? The purpose of this section is to show that earlier contributions in location theory have uncovered several of the main ideas used in NEG to answer the above questions. We do not intend to provide here an exhaustive survey of what had been accomplished by economists in pre-Krugman times but, more modestly, to show that the main ingredients had already been there, sometimes for a long period. This will allow us to assess why there is an 'N' in NEG.

2.1 The location of a firm

The location problem of a firm arises because some of its activities are indivisible. In Koopmans's (1957, p. 154) words:

> "without recognizing indivisibilities - in human person, in residences, plants, equipment, and in transportation - urban location problems, down to those of the smallest village, cannot be understood."

[5] In the chapter by Baldwin and Martin, the reader will find a discussion of agglomeration forces that are growth-specific.

More precisely, for the location problem of a firm to be not trivial, there must be some sort of increasing returns at the plant level as well as transport costs: increasing returns lead the firm to concentrate its production in a few plants, whereas transport costs raise the issue of where to locate these plants.

2.1.1 Increasing returns vs. transport costs

The fundamental trade-off of economic geography between increasing returns and mobility costs has been recognized by various scholars interested in the formation of human settlements. This should not come as a surprise because increasing returns and mobility costs may take quite different forms, thus making them applicable to a wide range of situations. One example is provided by the following statement made by W. M. Flatters Petrie in his *Social Life in Ancient Egypt* published in 1923 (pp. 3-4):

> "It has been noticed before how remarkably similar the distances are between the early nome capitals of the Delta (twenty-one miles on an average) and the early cities of Mesopotamia (averaging twenty miles apart). Some physical cause seems to limit the primitive rule in this way. Is it not the limit of central storage of grain, which is the essential form of early capital ? Supplies could be centralised up to ten miles away; beyond that the cost of transport made it better worthwhile to have a nearer centre."

This echoes Lösch (1940) who writes about two decades later:

> "We shall consider market areas that are not the result of any kind of natural or political inequalities but arise through the interplay of purely economic forces, some working toward concentration, and other toward dispersion. In the first group are the advantages of specialization and of large-scale production; in the second, those of shipping costs and of diversified production." (p. 105 of the English translation)

Observe that the same trade-off has been modeled by Kuehn and Hamburger (1963) in a planning context: given a spatial distribution of requirements for a particular commodity, fixed costs must be incurred for locating the facilities that produce this commodity and transport costs must be borne to ship it from the facilities to the consumers. The aim of the model is then to determine the number and locations of facilities so as to minimize the sum of production and transport costs.[6]

2.1.2 Weber and the location of the firm

The oldest formal analysis of the location of a firm is the minisum location problem, which consists in locating a plant in the plane. The aim is to minimize the weighted sum of Euclidean distances from that plant to a finite

[6] See, e.g. Hansen *et al.* (1987) for an economic-oriented survey of this model and of related ones.

number of sites corresponding to the markets where the plant purchases its inputs and sells its outputs; the weights represent the quantities of inputs and outputs bought and sold by the plant, multiplied by the appropriate freight rates (Weber, 1909). When the sites are not collinear, the minisum problem has no analytical solution and one must pertain to an algorithm to solve it (Weiszfeld, 1936). However, it is possible to derive special, but meaningful, results. A site is said to be a *dominant place* if its weight is greater than or equal to the sum of the weights of all the other sites. In this case, the dominant place is the optimal solution to the minisum problem (Witzgall, 1964). Such a simple result may explain the locational decision made by seemingly different firms to set up in a large metropolitan area as well as it allows to understand the location of steel mills next to the iron mines in the 19th century and later next to their main markets.

Since the 1960s, the focus in firm location theory has moved from the least transport cost approach to the more standard microeconomic approach of profit maximization. The corresponding integration of additional variables within the Weberian framework has permitted a better understanding of the influence of several economic and geographic factors in electing a location. This is exemplified by the work of Sakashita (1967) who demonstrates that, in the case of a segment connecting the market and the input source, the firm always chooses to set up at one of the two endpoints and never at an intermediate point. For that, freight rates have to be constant or to taper off as the distance covered increases, a condition that characterizes all modern transport technologies. Thus, models in firm location theory tell us something fundamental for the space-economy: *new firms have a high propensity to settle at places where economic activities are already established.*

2.2 The location of an industry

In location theory, the passage from a firm to an industry is not smooth because the locational decisions of several firms can hardly be studied within a perfectly competitive framework.

2.2.1 Starrett and the breakdown of the competitive price mechanism

The most elegant and general model of a competitive economy is the Arrow-Debreu model. The economy is formed by agents (firms and consumers) and by commodities (goods and services). A firm is characterized by a set of production plans, each production plan describing a possible input-output relation. A consumer is identified by a relation of preference, by a bundle of initial resources and by shares in the firms' profits. Roughly speaking, when firms' technologies exhibit nonincreasing returns to scale, there exist market prices (one per commodity), a production technology for each firm and a consumption bundle for each consumer that satisfy the following conditions: at the prevailing prices (i) supply equals demand for each commodity; (ii) each firm maximizes its profit subject to its possible production technologies;

and (iii) each consumer maximizes her utility under her budget constraint defined by the value of her initial endowment and her shares in firms' profits. In other words, all markets clear while each agent chooses her most preferred action at the equilibrium prices.

In a world à la Arrow-Debreu, a commodity is defined not only by its physical characteristics, but also by the place where it is available. The same good traded at different places is treated, therefore, as different commodities. Within this framework, choosing a location is part of choosing commodities. Hence, the Arrow-Debreu model aims at integrating spatial interdependence of markets into general equilibrium in the same way as other forms of interdependence. Unfortunately, the spatial impossibility theorem by Starrett (1978) shows that things are not that simple.

Suppose that production activities are not perfectly divisible and, hence, associate an *address* with any specific production activity. Furthermore, consider the extreme case of a *homogeneous* space with a finite number of locations. By a homogeneous space, we mean the following two conditions: (i) the production set of a firm is the same in all locations and (ii) consumers' preferences are the same at all locations. Such an assumption is made in order to control for the impact that first nature may have on the distribution of economic activity. This is because we are interested in finding economic mechanisms that explain agglomeration without appealing to physical attributes of locations (second nature). For the rest, the economy follows the lines of the competitive framework as described in the foregoing. Then, we have:

Theorem 1 *(***The Spatial Impossibility Theorem)**. *Consider an economy with a finite number of locations and a finite number of consumers and firms. If space is homogeneous, transport is costly and preferences are locally nonsatiated, then there is no competitive equilibrium involving transportation.*

What does it mean? If economic activities are perfectly divisible, a competitive equilibrium exists and is such that each location operates as an autarchy. For example, if consumers have the same preferences and identical initial endowments, regions have the same relative prices and the same production structure (backyard capitalism). This is hardly a surprising outcome since, by assumption, there is no reason for the economic agents to distinguish among locations and since each activity can operate at an arbitrarily small level. It is worth stressing, however, that the autarchic equilibrium does not necessarily involve a uniform distribution of activities.

Once economic activities are *not* perfectly divisible, as observed by Starrett (1978, p. 27), the transport of some goods between some places becomes unavoidable:

> "as long as there are some indivisibilities in the system (so that individual operations must take up space) then a sufficiently complicated set of interrelated activities will generate transport costs" (Starrett, 1978, p. 27)

In this case, the spatial impossibility theorem tells us that no competitive equilibrium with trade across locations exists. Since activities are not perfectly divisible (thus implying the presence of nonconvexities in the economy), firms and consumers may want to be separate because each must choose to use a positive amount of land that cannot be made arbitrarily small, even though the individual land consumption is endogenous. Physical separation and the homogeneity of space then place firms and consumers in a relation of symmetry such as the only spatial factor that matters to them is their position relative to the agents with whom they trade. In this context, *the fact that shipping goods is costly and that agents have an address generates nonconvexities in production or consumption sets that prevent the competitive setting from handling such a system of exchanges.* It is in that sense that the Arrow-Debreu framework fails to handle the spatial question: competitive equilibria, when they exist, involve no trade across locations.[7]

If we want to understand something about the spatial distribution of economic activities, especially the formation of major economic agglomerations, it follows from the spatial impossibility theorem that we must assume either that *space is heterogeneous* (as in the neoclassical theory of international trade or in land use models à la von Thünen), or that *externalities* exist and are many (as in modern urban economics), or that *markets are imperfect* (as in spatial competition theory or in economic geography). Although it is obvious that space is heterogenous, diversity of resources seems weak as a sole explanation for the existence of large metropolises as well as for the persistence of substantial regional income inequalities. For this reason, in what follows, we consider the other two explanations.

2.2.2 Cities and product variety

Whereas Marshallian externalities have a long standing tradition in modeling the supply side of a spatial economy (Henderson, 1988),[8] it is less known that they also operate on the demand side. In a very insightful - but not much quoted - paper, Haig (1926) argues that cities offer a great number of people a large assortment of consumption goods and services. He views the advantages associated with variety (*varietas delectat*) as being so large that the question is changed from "Why live in the city?" to "Why not live in the city?". In other words, migration to a large city may be explained purely by a consumption motive:

> "in a large city an individual may derive a higher utility from spending a given amount of income than in a small town ...even if the prices for commodities obtainable at both locations are higher in the former than in the latter." (Stahl, 1983, p. 318)

a result which follows directly from the convexity of preferences. What will then influence the development of an urban center is the cost of delivering

[7] For a more detailed discussion, see Fujita and Thisse (2002, ch. 2).

[8] See the chapter by Duranton and Puga in this volume for a survey of the microeconomic underpinnings of such externalities.

such assortments in the countryside. We thus have the main ingredients that govern the agglomeration and dispersion forces at work in NEG models. In addition, Haig observes that

> "The great bulk of population...must work and must consume most of what they earn where they earn it. With them consumption and production is practically a simultaneous process and must be carried on for the most part in the same place. To them location is of interest both in its effects upon production and in its effects upon consumption." (pp. 185-186)

This drives the formation of an agglomeration in the CP model: when workers move, they do so with their expenditures as well as with their labor.

Along the same lines, Lampard (1955) makes it clear that a variety-like argument similarly applies to intermediate goods used by firms because "the principal function of the city today in terms of employment it creates is the provision of services rather than manufactures."

2.2.3 Hotelling and spatial competition

Ever since the pioneering work of Hotelling (1929), it is generally accepted that competition for market areas is a centripetal force that would lead sellers to congregate, a result known as the Principle of Minimum Differentiation. The two ice-cream men problem provides a neat illustration of this principle. Two merchants selling the same ice-cream at the same fixed price compete in location for consumers who are uniformly distributed along a beach of length L; each consumer purchases one cone of ice-cream from the nearer firm. Since Lerner and Singer (1937), it is well known that the unique Nash equilibrium of this game is given by the location pair $s_1^* = s_2^* = L/2$. In words, the two ice-cream sellers choose to locate back to back at the market center. This is due to the so-called 'market area effect': each seller's profit is an increasing function of the fraction of the beach to which she has privileged access.

However, things become more complex when (mill) prices are brought into the picture. Hotelling considers a two-stage game where the firms first simultaneously choose their locations and afterwards their prices. The market equilibrium may then be viewed as the interplay between a dispersion force and an agglomeration force. To illustrate how this trade-off works, let $\pi_1^* = \pi_1(p_1^*, p_2^*, s_1, s_2)$ be firm 1's profit evaluated at the equilibrium prices $p_i^*(s_1, s_2)$ corresponding to the location pair $s_1 < s_2$. Then, since $\partial\pi_1^*/\partial p_1^* = 0$, we have:

$$\frac{d\pi_1^*}{ds_1} = \frac{\partial\pi_1}{\partial p_2}\frac{\partial p_2^*}{\partial s_1} + \frac{\partial\pi_1^*}{\partial s_1}$$

In general, the terms in the right hand side of this expression can be signed as follows. The first one corresponds to the *price effect* (the desire to relax price competition) and is expressed by the impact that a change in firm 1's

location has on price competition. Since goods are spatially differentiated, they are substitutes so that $\partial\pi_1/\partial p_2$ is positive; because goods become closer substitutes when s_1 increases, $\partial p_2^*/\partial s_1$ is negative. Hence the first term is negative: this is the dispersion force. The second term, which corresponds to the *market area effect*, is positive: this is the agglomeration force. Consequently, the impact of reducing the inter-firm distance upon firms' profits is a priori undetermined. Nevertheless, as established by d'Aspremont *et al.* (1979), if the transport costs are quadratic, a unique price equilibrium exists for any location pair and the two firms wish to set up at the endpoints of the market. In other words, *price competition is a strong centrifugal force.*

This dispersion of firms turns out to be very sensitive to a particular assumption of the model, namely firms sell an homogeneous good. Following Haig and NEG, we find it more realistic to assume that they sell differentiated products whereas consumers like product variety. Consequently, even if prices and locations do not vary, they do not always purchase from the same vendor. The idea that consumers distribute their purchases between several sellers is not new in economic geography and goes back at least to Reilly (1931) who formulated the so-called *gravity law of retailing.* Since individual demands are perfectly inelastic, such a behavior is naturally modeled by discrete choice theory. In the special case of the logit (McFadden, 1974), a consumer at x buys from firm i with a frequency given by

$$\mathbf{P}_i(x) \equiv \frac{\exp -(p_i + t\,|x - s_i|)/\upsilon}{\sum_{j=1}^{M} \exp -(p_j + t\,|x - s_j|)/\upsilon} \tag{1}$$

where t is the transport rate (transport costs are linear in distance) and υ the degree of product differentiation.[9] Let c be the common marginal production cost. Then, we have (de Palma *et al.*, 1985):

Proposition 2 (The Principle of Minimum Differentiation) *Assume that firms choose simultaneously their prices and locations. If consumers' purchasing behavior is described by (1) and if* $\upsilon/tL \geq 1/2$ *holds, then* $p_i^* = c + \upsilon n/(n-1)$ *and* $s_i^* = L/2$, $i = 1, ..., n$ *is a Nash equilibrium.*

Therefore, firms choose to agglomerate at the market center when their products are differentiated enough, when transport costs are low enough, or both.

Considering a more general setting in which the optimal behavior of a firm depends on what households and other firms do, while the optimal behavior of a household depends on what firms and other households do, Papageorgiou and Thisse (1985) describe the process of interaction between the two classes of agents as follows:

[9]It is worth noting here that the logit and the CES, which is extensively used in NEG, are closely related in that both models can be derived from the same distribution of consumer tastes; the only difference is that consumers buy one unit of the product in the former and a number of units inversely related to its price in the latter (Anderson *et al.*, 1992, chs. 3-4).

> "Households are attracted by places where the density of firms is high because opportunities there are more numerous, and they are repulsed by places where the density of households is high because they dislike congestion. Firms are attracted to places where the density of consumers is high because there the expected volume of business is large, and they are repulsed by places where the density of sellers is high because of the stronger competition prevailing there."(p. 20)

Those authors then show that, when varieties are sufficiently differentiated and/or transport costs are low enough, the interaction among firms and households leads to a spatial equilibrium in which both sellers and customers distribute themselves according to two bell-shapes curves that sustain each other. This confirms, within a broader framework, Hotelling's principle of minimum differentiation.

2.3 Where did we stand in 1990?

Putting all those things together, it follows that the legacy of location theory can be summarized in five points: (I) the economic space is the outcome of a trade-off between various forms of increasing returns and different types of mobility costs; (II) price competition, high transport costs and land use foster the dispersion of production and consumption; therefore (III) firms are likely to cluster within large metropolitan areas when they sell differentiated products and transport costs are low; (IV) cities provide a wide array of final goods and specialized labor markets that make them attractive to consumers/workers; and (V) agglomerations are the outcome of cumulative processes involving both the supply and demand sides.[10] Consequently, *the space-economy has to be understood as the outcome of the interplay between centripetal and centrifugal forces*, an idea put forward by geographers and regional scientists long ago, *within a general equilibrium framework accounting explicitly for market failures.*

Those five points capture also the main ingredients of NEG with a focus on pecuniary rather than technological externalities. Thus, there is little new about them. Nevertheless, before NEG they were not knitted together and were often available only in specialized studies. In addition, they were (at best) developed within partial equilibrium models. Therefore, *what was missing was a general equilibrium framework with imperfect competition connecting these various insights and allowing for a detailed study of their interactions.* Here lies the main contribution of NEG: "to combine old ingredients through a new recipe", even though Paul Krugman failed to acknowledge (and probably was not aware of) most of them when he published his *Journal of Political Economy* paper.

[10]The connections with early regional development theories, such as Myrdal (1957), have been made clear by Krugman himself from the very beginning of his work. This is why they have not been discussed here.

3 Where do firms locate: the home market effect

The first step to understand how NEG models work is to consider the home market effect (HME). In the case of a two-region economy, the HME implies that the location with larger local demand succeeds in attracting a more than proportionate share of firms in imperfectly competitive industries. This pattern of demand-driven specialization maps into trade flows and generates the theoretical prediction that large regions should be net exporters of goods produced under increasing returns and imperfect competition. The HME deals with the location of an industry when the spatial distribution of consumers is fixed, a topic investigated by Lösch (1940, ch. II) and extensively studied in location theory since then. There are differences, however. In standard location theory, space is represented by a one- or two-dimensional space (Eaton and Lipsey, 1977; Beckmann and Thisse, 1986); space is here described by two regions, as in trade theory. Even though poorer from a spatial point of view, the HME is nevertheless much richer in terms of microeconomic content. Having said that, the intuition behind the HME lies in standard location theory. Since a profit-maximizing firm also minimizes the transport costs it incurs when delivering its output, everything else equal it will locate in the larger market, which corresponds to a dominant place in the sense of Weber (see section 2.1.2). Nevertheless, not all firms will locate in the bigger market because positive transport costs allow firms to relax price competition by locating far from their competitors (see section 2.2.3). Hence some firms may want to set up in the smaller market.

The HME has been generally discussed under monopolistic competition.[11] On such a market structure are also generally based all existing models in NEG. For this reason, we choose to focus all that follows on monopolistic competition. However, we want to stress that, despite major differences in market structure, the HME bears some strong resemblance with the principle of minimum differentiation discussed in section 2.2.3. In both cases, consumers are dispersed and firms lure to the location with the highest potential for demand: the larger market in trade theory and the market center in spatial competition.

3.1 The market structure problem

In his review of Chamberlin's (1933) book, Kaldor (1935) claims that a firm affects the sales of its neighboring firms, but not distant ones. The impact of its price reduction is, therefore, not symmetric across all locations. In other words, there are good reasons to believe that competition across locations is inherently oligopolistic (Eaton and Lipsey, 1977; Gabszewicz and Thisse, 1986). Unfortunately, models of spatial competition are plagued by the frequent nonexistence of an equilibrium in pure strategies (Gabszewicz and Thisse, 1992). Thus, research has faced a modeling trade-off: to appeal to mixed strategies, or to use monopolistic competition in which interactions

[11] See Head *et al.* (2002) for a discussion of the HME under both monopolistic and oligopolistic competition.

between firms are weak. For the sake of simplicity, Krugman and most of the economics profession have retained the second option, which is not unreasonable once we address spatial issues at a macro-level. In addition, models of monopolistic competition have shown a rare ability to deal with a large variety of issues related to economic geography, which are otherwise unsatisfactorily treated by the competitive paradigm (Matsuyama, 1995). However, it should be kept in mind that spatial competition should not be missed at the micro-level.

In the Dixit and Stiglitz (1977) setting, *monopolistic competition* emerges as a market structure determined by consumers' heterogeneous tastes and firms' fixed requirements for limited productive resources. On the demand side, the set of consumers with different tastes are aggregated into a representative consumer whose preferences exhibit *love for variety*: her utility is an increasing function not only of the amount of each variety of a horizontally differentiated good, but also of the total number of varieties available.[12] On the supply side, production exhibits economies of scale within varieties but no economies of scope across varieties, thus implying a one-to-one relationship between firms and varieties. Consequently, each firm supplies one and only one variety (monopolistic). However, there are no entry or exit barriers so that prices are just enough to cover average cost (competition). Finally, firms are so many that they do not interact directly but only indirectly through aggregate demand effects. Formally, we assume that there is a *continuum of firms*.

The continuum approach does not imply that firms' behavior is totally nonstrategic. Indeed, each firm must figure out what will be the total output (or, alternatively, the average price index) in equilibrium when choosing its own quantity or price, or when deciding whether to enter the market. This is not what we encounter in a differentiated oligopolistic market when individual decisions made by competitors are needed by each firm. Here, we have a setting in which each firm must know only a global statistics about the market but not its details. We believe that using a statistics of the market is a particularly appealing way to capture the idea of monopolistic competition because it saves the essence of competition by forcing each firm to account for the aggregate behavior of its competitors.

Furthermore, the continuum assumption is probably the most natural way to capture Chamberlin's intuition regarding the working of a 'large group' industry, while allowing us to get rid of the 'integer problem' which often leads to inelegant results and cumbersome developments. Note also that, unlike oligopoly theory, which is plagued by the differences between the Bertrand and Cournot settings, the distinction between price competition and quantity competition becomes immaterial in monopolistic competition. Indeed, being negligible to the market, each firm behaves as a monopolist on the residual demand, which makes it indifferent between using price or quantity as a strategy. Last, this modeling strategy allows one to respect the

[12]The CES utility used by Dixit and Stiglitz (1977) is indeed an aggregate for a particular population of heterogeneous consumers (Anderson *et al.,* 1992, chs. 3-4).

indivisibility of an agent's location (her 'address') while avoiding to appeal to the existence of strong nonconvexities associated with large agents. At the same time, it leads to a description of the regional shares of economic and demographic magnitudes by means of continuous variables.

Although we consider only specific models of monopolistic competition such as the CES and the linear models, we expect the results obtained in these two different settings to be representative of general tendencies.

3.2 The basic framework

We consider a $2 \times 2 \times 2$ setting. The economic space is made of two regions (A and B). The economy has two sectors, the 'modern' sector X and the 'traditional' sector Z. There are two production factors, capital (H) and labor (L).

For the HME to arise, the two regions must differ in their expenditures. This could happen for various reasons and we find it convenient to consider a simple setting. The economy is endowed with H capitalists and L workers each supplying one unit of their corresponding factor inelastically. Capital is perfectly mobile between regions, whereas labor is immobile. Specifically, workers are distributed so that a fraction $\theta \in (0,1)$ resides in region A: $L_A = \theta L$. Without loss of generality, that region is assumed to host a larger number of workers ($\theta > 1/2$). To rule out comparative advantage à la Heckscher-Ohlin, capitalists are distributed according to the same fraction θ in region A: $H_A = \theta H$. Relative factor endowments are then the same across regions.

The Z-sector produces a homogeneous good under constant returns to scale and perfect competition. In particular, one unit of output requires one unit of L. Profits maximization then yields $p_i^Z = w_i$ where w_i is the wage. This good is costlessly traded between regions so that its price is the same everywhere: $p_A^Z = p_B^Z$. It is convenient to choose the homogeneous good as the numéraire, implying $p_A^Z = p_B^Z = 1$ and $w_A = w_B = 1$.

The X-sector produces a continuum of horizontally differentiated varieties under increasing returns. Each variety is supplied according to the same increasing-return technology: the production of $x(s)$ units of variety s requires a fixed amount f of capital and a variable amount $mx(s)$ of labor so that the total cost of the firm producing variety s is given by

$$TC_i(s) = r_i f + w_i m x_i(s) \tag{2}$$

where r_i is the rental rate of capital in region i. Because there are increasing returns to scale but no scope economies, each variety is produced by a single firm.[13] Indeed, since consumers have a preference for variety, any firm obtains a higher share of the market by producing a differentiated variety than by replicating an existing one. Furthermore, shipping a variety across re-

[13]This assumption agrees with what is called the principle of differentiation in industrial organization.

gions is costly but intraregional transport costs are zero.[14] We assume that regional markets are *segmented*: each firm sets a delivered price specific to the market in which its variety is sold.[15] In the first model discussed below (3.2.1), demands have the same elasticity across locations so that both mill and discriminatory pricing policies yield the same equilibrium prices and outputs. This equivalence no longer holds in the second model (3.2.2).

The market equilibrium is the outcome of the interplay between a dispersion force and an agglomeration force, very much as the spatial competition model à la Hotelling. The centripetal force lies in *market access*: the transport cost saving provided by locating in the larger region, the counterpart of the market area effect in spatial competition. The centrifugal force lies in *market crowding*: the fiercer competition that arises when firms locate back to back, which corresponds to the price effect in spatial competition. Hence, the forces are the same, even though we consider two locations here, whereas there is a continuum of locations in spatial competition. The main difference is that we use here a general equilibrium framework instead of a partial equilibrium model.

Let $\lambda \in (0,1)$ be the fraction of capital employed in region A so that $(\theta - \lambda)H > 0$ (< 0) measures the extent of capital flows into (out of) A. Denote by $r_i(\lambda)$ the rental rate of capital in region $i = A, B$ when its spatial distribution is $(\lambda, 1-\lambda)$. A *spatial equilibrium* arises at $\lambda \in (0,1)$ when

$$\Delta r(\lambda) \equiv r_A(\lambda) - r_B(\lambda) = 0$$

or at $\lambda = 0$ when $\Delta r(0) \leq 0$, or at $\lambda = 1$ when $\Delta r(1) \geq 0$. Such an equilibrium always exists when $r_i(\lambda)$ is a continuous function of λ (Ginsburgh *et al.*, 1986).

In the absence of a general model of monopolistic competition, we add one more 2 to the $2 \times 2 \times 2$ setting by discussing the two specific models that have been used so far in NEG.[16]

3.2.1 A nonlinear model with fixed mark-ups: CES utility and iceberg transport costs

The preference ordering of a consumer living in region i is captured by the utility function:

$$U_i = Q_i^{\mu} Z_i^{1-\mu} \tag{3}$$

[14]This shows once more the trade-theory origin of most NEG models. Intraregional transport costs should also be positive but different from the interregional transport costs. By changing the value of these costs, one may study the impact of the quality of transport infrastructures on the distribution of activities. See Martin and Rogers (1995) for a first attempt along these lines.

[15]There are many good reasons to believe that firms want to use spatial separation to segment their market (Thisse and Vives, 1988), wheras empirical work confirms the assumption that international, or even interregional, markets are still very segmented (Greenhut, 1981; Head and Mayer, 2000).

[16]In the HME context, they have been developed respectively by Martin and Rogers (1995) and Ottaviano (2001a).

where

$$Q_i = \left(\int_0^N q_i(s)^{\frac{\sigma-1}{\sigma}} ds \right)^{\frac{\sigma}{\sigma-1}} \tag{4}$$

is the consumption of good X, Z_i the consumption of the numéraire, $q_i(s)$ the consumption of variety s of good X, and N the total number (mass) of varieties. Because each variety is negligible, $\sigma > 1$ is both the elasticity of demand of any variety and the elasticity of substitution between any two varieties.

Standard utility maximization of (3) yields CES demand by residents in region i for a variety produced in location j:

$$q_{ji}(s) = \frac{p_{ji}(s)^{-\sigma}}{P_i^{1-\sigma}} \mu Y_i \tag{5}$$

where p_{ji} is the consumer price of a variety produced in j and sold in i, P_i is the local CES price index associated with (4):

$$P_i = \left[\int_{s \in n_i} p_{ii}(s)^{1-\sigma} ds + \int_{s \in n_j} p_{ji}(s)^{1-\sigma} ds \right]^{\frac{1}{1-\sigma}} \tag{6}$$

where n_i is both the set and the number of varieties produced in region i so that $n_A + n_B = N$. The regional income Y_i consists of capital rental rates (R_i) and wages ($w_i L_i$):

$$Y_i = R_i + w_i L_i. \tag{7}$$

Thus, the representative consumer in region i maximizes utility (3) subject to the following budget constraint:

$$\int_{s \in n_i} p_{ii}(s) q_{ii}(s) ds + \int_{s \in n_j} p_{ji}(s) q_{ji}(s) ds + p_i^Z Z_i = Y_i. \tag{8}$$

Moreover, as already discussed, costlessly trade and perfect competition in sector Z together with the choice of numéraire imply $p_i^Z = w_i = 1$.[17] Trade in X, on the contrary, is inhibited by frictional trade barriers, which are modeled as iceberg costs à la Samuelson: for one unit of the differentiated good to reach the other region, $\tau \in [1, \infty)$ units must be shipped.[18]

Due to the fixed input requirement f, capital market clearing implies that, in equilibrium, the number of firms is determined by $N = H/f$ with

$$n_A = \frac{\lambda H}{f} \qquad n_B = \frac{(1-\lambda)H}{f} \tag{9}$$

[17] Wage equalization holds as long as the homogeneous good is produced in both regions. That is the case when a single region alone cannot supply the economy-wide demand, i.e. when good Z has a large weight in utility (μ small) and product variety is highly valued by consumers (σ small). The exact condition is $\mu < (1-\theta)/(1-\theta/\sigma)$, which is assumed to hold from now on.

[18] Hence, transport costs have the nature of an *ad valorem* sales tax equal to $\Upsilon - 1$. We do not know any version of the present model in which transport costs would be additive.

so that *the number of active firms in a region is proportional to the amount of capital that is employed locally.*

Using (2) a typical firm located in region i maximizes profit:

$$\Pi_i(s) = p_{ii}(s)q_{ii}(s) + p_{ij}(s)q_{ij}(s) - m[q_{ii}(s) + \tau q_{ij}(s)] - r_i f$$

where, due to the choice of numéraire, the wage w_i is set equal to 1, while $\tau q_{ij}(s)$ represents total supply to the distant location j inclusive of the fraction of variety i that melts on the way. The first order condition for profit maximization gives:

$$p_{ii}^*(s) = m\sigma/(\sigma - 1) \qquad \text{and} \qquad p_{ij}^*(s) = \tau m\sigma/(\sigma - 1) \tag{10}$$

for every i and j.[19] Using (10), the CES price index (6) simplifies to:

$$P_i = \frac{m\sigma}{\sigma - 1}(n_i + \phi n_j)^{\frac{1}{1-\sigma}} \tag{11}$$

where $\phi \equiv \tau^{1-\sigma} \in (0, 1]$ is a measure of the *freeness of trade*, which increases as τ falls and is equal to one when trade is costless ($\tau = 1$). Since the total number of X-firms is given by $N = H/f$, the price index (11) decreases (increases) with the number of local (distant) firms.

Due to free entry and exit, there are no profits in equilibrium. This implies that a firm's scale of production is such that operating profits exactly match the fixed cost paid in terms of capital. In other words, the equilibrium rental rate corresponding to (9) is determined by a bidding process for capital, which ends when no firm can earn a strictly positive profit at the equilibrium market prices. That is, a firm's operating profits are entirely absorbed by the cost of capital:

$$r_i f = p_{ii}^*(s)q_{ii}(s) + p_{ij}^*(s)q_{ij}(s) - m[q_{ii}(s) + \tau q_{ij}(s)]$$

which, given (5), (10) and (11), yields

$$r_i = \frac{mx_i}{f(\sigma - 1)} \tag{12}$$

where $x_i = q_{ii}(s) + \tau q_{ij}(s)$ is the total production by a typical firm in location i.

Market clearing for a typical variety produced in region i implies:

$$x_i^* = \frac{\sigma - 1}{m\sigma}\left(\frac{\mu Y_i}{n_i + \phi n_j} + \frac{\phi \mu Y_j}{\phi n_i + n_j}\right). \tag{13}$$

Using (9) and (12), we may rewrite the rental rate in region A as follows:

$$r_A(\lambda) = \frac{\mu}{\sigma H}\left[\frac{Y_A}{\lambda + \phi(1 - \lambda)} + \frac{\phi Y_B}{\phi\lambda + (1 - \lambda)}\right] \tag{14}$$

[19] Note the difference with Krugman: equilibrium prices depend on the cost of the mobile factor. Here, they are constant because the cost of the immobile factor is given.

with a symmetric expression holding for region B.

For $\lambda \in (0,1)$ the equilibrium distribution of firms solves $r_A(\lambda) = r_B(\lambda) = r$, which implies:

$$Y_A = \theta(r\,H + L) \qquad Y_B = (1-\theta)(r\,H + L). \tag{15}$$

Plugging (15) into (14) and solving the two resulting equations for λ and r gives the (interior) equilibrium distribution of firms:

$$\lambda^* = \frac{1}{2} + \frac{1+\phi}{1-\phi}\left(\theta - \frac{1}{2}\right) > \theta > \frac{1}{2}. \tag{16}$$

This expression reveals the presence of the HME: *the larger region hosts a more than proportionate share of firms.* In addition, this effect is accelerated as θ rises because $d\lambda^*/d\theta > 1$. Finally, λ^* rises continuously with ϕ so that the agglomeration process is smooth with respect to falling transport costs.

For ease of interpretation, (16) may be rewritten as follows:

$$(1+\phi)(\theta - 1/2) - (1-\phi)(\lambda^* - 1/2) = 0. \tag{17}$$

¿From left to right, the first term depends on the spatial distribution of consumers θ. Since the coefficient of $(\theta - 1/2)$ is positive, this term measures the *market access* advantage of the larger region in the presence of trade barriers. The second term in (17) depends on the international distribution of firms λ. Since the coefficient of $(\lambda^* - 1/2)$ is negative, that term measures the *market crowding* disadvantage of the region that hosts the larger number of firms. Lower trade barriers (τ) and a smaller elasticity of substitution (σ) make ϕ larger, thus strengthening the access advantage while weakening the crowding disadvantage. In particular, we have $d^2\lambda^*/d\theta\,d\phi > 0$, which reveals the magnification of the HME by freer trade.

While enlightening, the CES set-up with iceberg costs faces some shortcomings. In particular, it yields a demand system in which the own-price elasticities of demands are constant, identical to the elasticities of substitutions and equal to each other across all varieties. This entails that the equilibrium mark-up is independent of the spatial distribution of firms and consumers. Though convenient from the analytical point of view, such a result conflicts with research in spatial price theory where it is shown that demand elasticity varies with distance while profits change with the level of demand and the intensity of competition (Greenhut *et al.*, 1987). Moreover, the iceberg assumption also implies that any increase in the price of the shipped good is accompanied by a proportional increase in its transport cost, which is unrealistic. All this entangles the economic meanings of the various parameters, thus leading to unclear comparative static results.

3.2.2 A linear model with variable mark-ups: quadratic utility and linear transport costs

To avoid some of the pitfalls of the CES and iceberg, we now consider a model that differs in the specification of preferences and transport costs.

Preferences are described by the following quasi-linear quadratic utility:

$$U_i = \alpha \int_0^N q_i(s)ds - \frac{\beta-\gamma}{2}\int_0^N [q_i(s)]^2 ds - \frac{\gamma}{2}\left[\int_0^N q_i(s)ds\right]^2 + Z_i. \tag{18}$$

The parameters in (18) are such that $\alpha > 0$ and $\beta > \gamma > 0$. In this expression, α expresses the intensity of preferences for the differentiated product, whereas $\beta > \gamma$ means that consumers are biased toward a dispersed consumption of varieties. In particular, the quadratic utility function exhibits love for variety as long as $\beta > \gamma$. Finally, for a given value of β, the parameter γ expresses the substitutability between varieties: the higher γ, the closer substitutes the varieties. In the limit, when $\beta = \gamma$, (18) degenerates into a standard quadratic utility defined over a homogenous product.

Standard utility maximization under (8) yields linear demands by residents in region i for a variety produced in location j:

$$q_{ji}(s) = [a - (b + cN)\, p_{ji}(s) + cP_i]^+ \tag{19}$$

where $[f]^+$ denotes the positive part of f, p_{ji} the consumer price of a variety produced in j and sold in i,

$$P_i \equiv n_i\ p_{ii} + n_j\ p_{ji}$$

and

$$\begin{aligned} a &\equiv \frac{\alpha}{\beta + (N-1)\gamma} \qquad b \equiv \frac{1}{\beta + (N-1)\gamma} \\ c &\equiv \frac{\delta}{(\beta-\gamma)[\beta + (N-1)\gamma]}. \end{aligned}$$

Clearly, P_i/N can be interpreted as the price index prevailing in region i. Thus, (19) encapsulates the idea that the demand of a certain variety falls when its own price rises not only in absolute terms (*own price effect*) but also relatively to the average price (*differential price effect*), which seems to be the essence of monopolistic competition.

On the supply side, the only difference with respect to the foregoing is that transport costs are not of the iceberg type. Specifically, the varieties of the modern sector are traded at a cost of $t > 0$ units of the numéraire per unit shipped between the two regions.[20] Accordingly, a representative firm in i maximizes its profits, which, after using (19), are defined by:

$$\begin{aligned} \Pi_i(s) = [p_{ii}(s) - m]\ [a - (b + cN)\, p_{ii}(s) + cP_i]\ M_i + \\ [p_{ij}(s) - m - t]\ [a - (b + cN)\, p_{ij}(s) + cP_j]\ M_j - r_i f \end{aligned} \tag{20}$$

[20]This means that transport costs have the nature of a specific sales tax. The implications of an iceberg-type transport cost have not been explored under linear demands.

where

$$M_A = \theta(L+H) \qquad M_B = (1-\theta)(L+H)$$

are the numbers of consumers in regions A and B respectively.

Market prices are obtained by maximizing profits while the rental rates of capital are determined as described above by equating the resulting profits to zero. Since we have a continuum of firms, each one is negligible in the sense that its action has no impact on the market. Hence, when choosing its prices, a firm in A accurately neglects the impact of its decision over the two price indices P_A and P_B. In addition, because firms sell differentiated varieties, each one has some monopoly power in that it faces a demand function with finite and variable elasticity. On the other hand, since the price index enters the demand function (see (19)), a firm must account for the distribution of the firms' prices through some aggregate statistics, given here by the average market price, in order to find its equilibrium price. As a consequence, the market solution is given by a Nash equilibrium in which each firm neglects its impact on the market but is aware that the market as a whole has a nonnegligible impact on its behavior.

Solving the first order conditions for profit maximization yields the equilibrium prices:

$$p_{ii}^* = \frac{1}{2}\frac{2[a+(b+cN)m]+n_j ct}{2b+cN} = \frac{P_i}{N} - \frac{t}{2}\frac{n_j}{N} \tag{21}$$

$$p_{ij}^* = p_{jj}^* + \frac{t}{2} = \frac{P_j}{N} + \frac{t}{2}\frac{n_i}{N} \tag{22}$$

which depend on the total number of active firms as well as on their distribution between the two regions. Due to trade barriers, in both regions domestic firms price below the average prices, P_A/N and P_B/N, the more so the relative number of foreign competitors; foreign firms price above average prices, the more so the larger the number of domestic competitors.

Substracting m and t from (22), we see that firms' prices net of transport costs are positive regardless of their spatial distribution if and only if

$$t < t_{trade} \equiv \frac{2(a-bm)}{2b+cN}. \tag{23}$$

The same condition must hold for consumers in B (A) to buy from firms in A (B), i.e. for the demand (19) evaluated at the equilibrium prices (21) and (22) to be positive for all λ. From now on, condition (23) is assumed to hold.

Using (23), we observe that more firms in the economy lead to lower market prices for the same spatial distribution $(\lambda, 1-\lambda)$ because there is more competition in each local market. Similarly, both the prices charged by local and foreign firms fall when the mass of local firms increases because competition is fiercer. Equilibrium prices also rise when the degree of product differentiation, inversely measured by c, increases provided that (23) holds.

The rental rate of capital prevailing in region A can be obtained by evaluating (20) at equilibrium prices, which yields the following quadratic expression in λ:

$$r_A(\lambda) = \frac{b+cN}{f}\left[(p_{AA}^* - m)^2\theta + (p_{AB}^* - m - t)^2(1-\theta)\right](L+H) \tag{24}$$

with a symmetric expression holding in region B.

Again, for $\lambda \in (0,1)$ the equilibrium distribution of firms solves $r_A(\lambda) = r_B(\lambda)$ so that, by (21), (22), and (24), it equals:

$$\lambda^* = \frac{1}{2} + \frac{2f(2a-2bm-bt)}{ctH}\left(\theta - \frac{1}{2}\right) > \theta > \frac{1}{2} \tag{25}$$

when (23) is satisfied, thus revealing the presence of the HME.[21] To gain insight about this result, we rewrite (25) as follows:

$$2(2a-2bm-bt)(\theta - 1/2) - ctN(\lambda^* - 1/2) = 0$$

which shows that in equilibrium the distribution of firms is again determined by the interaction of two terms. As before, since the coefficient of $(\theta - 1/2)$ is positive in so far as (23) holds, the first term measures the *market access* advantage of the larger region. Since the coefficient of $(\lambda^* - 1/2)$ is negative, the second term measures the *market crowding* disadvantage of the region that hosts a larger number of firms, an effect that does not appear in (17).

The quasi-linear quadratic set-up allows for clearer comparative statics results showing that the importance of the access advantage with respect to the crowding disadvantage grows as the own (b) and differential (c) price effects fall. In addition, more product differentiation (a smaller c) decreases the weight of market crowding. In the limit case of monopoly ($c = 0$), only market access considerations matter since a firm's operating profits are independent from other firms' locations (as in the single firm case considered by Sakashita, 1967). The relevance of market crowding also falls as the level of fixed costs grows, that is, as the number of competing firms N decreases. Finally, transport costs t affects both market access and crowding. In particular, lower t strengthens the former and weakens the latter. The reason why is that, with lower transport costs, a larger fraction of a firm's operating profits is independent of the location of competitors. In particular, we have $d^2\lambda^*/d\theta\, dt < 0$, which again shows the 'magnification' of the HME by freer trade.

The results of section 3.2 may be summarized as follows.

Proposition 3 (The Home Market Effect) *Consider an economy with two regions and two sectors. If the two regions differ only in terms of their expenditures on the modern good, then the market equilibrium involves a more than proportionate share of the modern sector in the region with larger expenditures.*

[21] Likewise, we have $d\lambda^*/d\theta > 1$

Accordingly, small permanent shocks to relative demands give rise to large permanent differences in regional sectoral specialization. This property cannot be readily extended to multi-regional set-ups. The three-region model in Krugman (1993) can be used to argue that the HME has no straightforward definition because there is no obvious benchmark against which to measure the 'more than proportionate' presence of imperfectly competitive firms. However, because the HME extends the idea of a dominant place in the Weber problem, existing results in location theory (Beckmann and Thisse, 1986) and 'classical' economic geography (Thomas, 2002) suggest, in the multiregional case, the existence of a hierarchy of regional markets, which depends on both the *size* of these markets but also their *relative position* within the space-economy.[22] Extensions along these lines would be a worthy contributions to the state of the art.

4 The core-periphery structure

The models discussed in the foregoing section reveal that imperfect competition and increasing returns can exacerbate exogenous differences in market size. More precisely, we have shown that small permanent shocks can give rise to large permanent differences between regions. In the present section, we discuss a framework, namely the core-periphery model (henceforth, CP model) originally due to Krugman (1991), which allows even small temporary shocks to cause large permanent differences between regions.

Because this happens when trade barriers are low, the proposed framework sheds light on the spatial aspects of the industrialization process, which may be collapsed in the following two steps described by Pollard (1981).[23] First, the symmetric pattern provides a fairly good approximation of the early configurations prevailing in Europe before the Industrial Revolution:

> "Before the industrial revolution, the gaps between different parts of Europe were much smaller than were to become later and some industrial activity not unlike that in Inner Europe was to be found almost everywhere." (p. 201)

Second, the formation of a core-periphery pattern seems to be a fair description of the way industrialization developed across regions:

> "the industrial regions took from them [their agricultural neighbors] some of their most active and adaptable labour and they encouraged them to specialize in the supply of agricultural produce, sometimes at the expense of some pre-existing industry running the risk thereby that

[22] When intraregional costs are taken into account, it is reasonable to assume that the larger (and richer) region has better infrastructures than the other. This amplifies the HME because higher intraregional costs amount to reducing the size of the local market (Martin and Rogers, 1995).

[23] The work of Tirado *et al.* (2002) about the impact on industrial location of economic integration in 19th century Spain also seems to provide a good illustration of the previsions made by CP models.

this specialization would permanently divert the colonized areas from becoming industrial themselves." (p. 11)

4.1 The basic framework

The setting departs from the one of the previous section under one major respect: H is not physical capital anymore but rather human capital that is embodied in workers. The crucial implication is that, differently from the foregoing, *a capital-owner can now offer her services only in the region where she resides*. Accordingly, H and L represent two types of workers that we call skilled and unskilled (*labor dualism*). The two types of workers differ in terms of their geographical mobility: the skilled workers are mobile between regions, whereas the unskilled workers are immobile.[24] Moreover, the unskilled workers are equally distributed between the two regions because we want regions to be a priori symmetric.[25]

The market equilibrium is again the outcome of the interplay between a market crowding effect and a market access effect. However, differently from the HME set-up, the mobility of workers affects both the supply and demand sides of the region of destination (and not just the supply side), thus making the size of the local market endogenous. The reason is that firms' relocation has to be matched by skilled migration so that the two may reinforce each other. This in turn induces some skilled living in the other region to move toward the region with more firms in which they may enjoy a higher standard of living. The resulting increase in the numbers of consumers creates a larger demand for the differentiated good which, therefore, leads additional firms to locate in this region. This implies the availability of more varieties in the region in question but less in the other because of scale economies at the firm's level. Consequently, as noticed by Krugman (1991, p. 486), there is *cumulative causation* à la Myrdal because these two effects reinforce each other: "manufactures production will tend to concentrate where there is a large market, but the market will be large where manufactures production is concentrated".[26]

Formally, (9) implies that λ now measures both the fraction of firms and the fraction of skilled workers in region A. Then, denoting by $v_i(\lambda)$ the indirect utility a skilled worker enjoys in region i, a *spatial equilibrium* arises at $\lambda \in (0,1)$ when

$$\Delta v(\lambda) \equiv v_A(\lambda) - v_B(\lambda) = 0$$

or at $\lambda = 0$ when $\Delta v(0) \leq 0$, or at $\lambda = 1$ when $\Delta v(1) \geq 0$. Such an equilibrium always exists when $v_r(\lambda)$ is a continuous function of λ (Ginsburgh *et*

[24] This extreme assumption is justified because the skilled are more mobile than the unskilled over long distances (SOPEMI, 1998).

[25] Note that Puga (1999) and Tabuchi and Thisse (2002a) deal with generalizations of the CP model in which all workers are mobile, whereas land is the only immobile factor. Such settings seem to be appropriate as an alternative foundation for studying urbanization.

[26] Observe that similar processes have been studied in spatial competition theory when consumers are imperfectly informed about the characteristics of the varieties sold by firms (Stahl, 1987).

al., 1986). However, this equilibrium is not necessarily unique. Stability is then used as a refinement to eliminate some of the equilibria.

The stability of an equilibrium is studied with respect to the following equation of motion:

$$\frac{d\lambda}{dt} = \begin{cases} \Delta v(\lambda) & \text{if } 0 < \lambda < 1 \\ \min\{0, \Delta v(\lambda)\} & \text{if } \lambda = 1 \\ \max\{0, \Delta v(\lambda)\} & \text{if } \lambda = 0 \end{cases} \tag{26}$$

Specifically, if $\Delta v(\lambda)$ is positive and $\lambda \in (0,1)$, workers move from B to A; if it is negative, they go in the opposite direction. Clearly, any spatial equilibrium is a steady-state of the process (26).

A spatial equilibrium is (locally) *stable* if, for any marginal deviation of the population distribution from the equilibrium, the equation of motion above brings the distribution of skilled workers back to the original one. When some skilled workers move from one region to the other, we assume that local labor markets adjust instantaneously. More precisely, the number of firms in each region must be such that the labor market clearing conditions (9) remain valid for the new distribution of workers. Wages are then adjusted for each firm to earn zero profits.

4.1.1 CES utility and iceberg transport costs

In Krugman (1991), utility is given by (3). The corresponding indirect utility differential is:

$$\Delta v(\lambda, \phi) \equiv \mu^{\mu}(1-\mu)^{1-\mu} \left\{ \frac{w_A(\lambda,\phi)}{[P_A(\lambda,\phi)]^{\mu}} - \frac{w_B(\lambda,\phi)}{[P_B(\lambda,\phi)]^{\mu}} \right\} \tag{27}$$

where $w_i(\lambda,\phi)$ is the wage prevailing in region $i = A, B$.[27] Substituting (9) into (11), the price index in region A is as follows:

$$P_A(\lambda,\phi) = \frac{m\sigma}{\sigma-1} \left[\lambda + \phi(1-\lambda)\right]^{\frac{1}{1-\sigma}} \left(\frac{H}{f}\right)^{\frac{1}{1-\sigma}} \tag{28}$$

with a symmetric expression holding for P_B. The presence of P_A and P_B in (27) adds a *new* item to the list of location effects. In particular, (28) shows that, for a given wage, the region with more skilled workers, and thus

[27] In Krugman (1991) skilled labor is used for both the fixed and variable costs of good X. For expositional purposes, we present here the analytically solvable version of his model put forth by Forslid and Ottaviano (2003), where only the fixed cost is incurred in terms of skilled labor whereas the variable cost is paid in terms of unskilled labor. Although the modified version and the original model exhibit the same qualitative behavior, there are some quantitative differences. The break and sustain points are larger while the no-black-hole condition is less stringent in the former than in the latter. In particular to pass from the modified version to the original model, one has to multiply μ by σ wherever it appears (see Fujita *et al.*, 1999). All together these properties imply that agglomeration forces are weaker in the modified version. The reason is weaker demand linkages: in Krugman (1991), the expenditures of a skilled worker are equal to the total revenues of the corresponding firm, in Forslid and Ottaviano (2003) they are equal to its operating profits only, which are indeed a fraction $1/\sigma$ of firm revenues.

more manufacturing firms, grants higher purchasing power, that is, higher consumer surplus. The reason is its lower price index as the larger number of domestic firms implies that fewer manufacturing varieties are imported and burdened by transport costs (*cost-of-living effect*). Therefore, *this additional effect teams up with the market size effect to support the agglomeration of manufactures against the market crowding effect.*

For the determination of skilled wages w_A and w_B, notice that the definition of incomes has changed to:

$$Y_A(\lambda,\phi)=\frac{L}{2}+w_A(\lambda,\phi)\,\lambda\,H \qquad Y_B(\lambda,\phi)=\frac{L}{2}+w_B(\lambda,\phi)\,(1-\lambda)\,H. \tag{29}$$

Plugging (29) into (14) and (15) and solving the two resulting equations together for w_A and w_B gives the equilibrium skilled wages:

$$w_i^*=\frac{\mu/\sigma}{1-\mu/\sigma}\frac{L}{2}\frac{2\phi\,n_i+[1-\mu/\sigma+(1+\mu/\sigma)\phi^2]\,n_j}{\phi\,(n_i^2+n_j^2)+[1-\mu/\sigma+(1+\mu/\sigma)\phi^2]\,n_in_j} \tag{30}$$

which, by (9), can be rewritten as a function of the distribution of firms λ and the freeness of trade ϕ. In the case of region A, this yields

$$w_A^*(\lambda,\phi)=\frac{1}{2}\frac{\mu/\sigma}{1-\mu/\sigma}\frac{L}{H}\frac{2\phi\,\lambda+[1-\mu/\sigma+(1+\mu/\sigma)\phi^2]\,(1-\lambda)}{\phi[\lambda^2+(1-\lambda)^2]+[1-\mu/\sigma+(1+\mu/\sigma)\phi^2]\lambda(1-\lambda)}$$

with a symmetric expression holding for $w_B(\lambda,\phi)$. Then we have:

$$\frac{w_A^*(\lambda,\phi)}{w_B^*(\lambda,\phi)}=\frac{2\phi\,\lambda+[1-\mu/\sigma+(1+\mu/\sigma)\phi^2]\,(1-\lambda)}{2\phi\,(1-\lambda)+[1-\mu/\sigma+(1+\mu/\sigma)\phi^2]\,\lambda}. \tag{31}$$

Differentiating (31) with respect to λ shows that the region with more workers offers them a higher (lower) wage whenever ϕ is larger (smaller) than the threshold:

$$\phi_r\equiv\frac{1-\mu/\sigma}{1+\mu/\sigma} \tag{32}$$

with $\phi_r\in(0,1)$.

This is the result of a trade-off between two opposing forces. On the one hand, for given transport costs, a larger number of skilled workers in a certain region entails a larger number of competing manufacturing firms. For given expenditures on manufactures, this depresses the local price index inducing a fall in local demand per firm (market crowding effect). Lower demand leads to lower operating profits and, therefore, lower skilled wages.[28] On

[28]While the focus here is on product market imperfections, factor market considerations may also be relevant. For example, Picard and Toulemonde (2003) introduce unions in the modern sector. They show that wage bargaining at the regional level acts as a dispersion force. By contrast, national bargaining destroys the centrifugal force generated by a nominal wage differential (Faini, 1999).

the other hand, hosting more firms also implies additional operating profits and thus additional skilled income, a larger fraction of which is spent on local manufactures. Accordingly, local expenditures are larger, which, for a given price index, increases demand per firm (market size effect). The former (latter) dominates the latter (former) whenever ϕ is smaller (larger) than ϕ_r. Note that this result is true for each λ.

All this is nicely captured by (32). Such an expression shows that *the market crowding effect is strong when transport costs (τ) are high* because firms sell mainly in their domestic market protected by high cost barriers. This effect is also strong when the own and cross price elasticity of demand for manufactures (σ) is large because a firm demand is quite sensitive to the price index. Finally, as intuition would have it, (32) also shows that *the market size effect is strong when the fraction of income spent on manufactures (μ) is large*. Thus, skilled wages are higher in the region with more skilled workers for small τ, large μ, and small σ.

Substituting (28) and (31) in (27) we obtain:

$$\Delta v(\lambda,\phi)=\frac{C}{\phi[\lambda^2+(1-\lambda)^2]+[1-\mu/\sigma+(1+\mu/\sigma)\phi^2]\lambda(1-\lambda)}\cdot\Delta V(\lambda,\phi)$$

where $C>0$ is a bundle of parameters independent of ϕ and

$$\begin{aligned}\Delta V(\lambda,\phi)\equiv&\frac{2\phi\,\lambda+[1-\mu/\sigma+(1+\mu/\sigma)\phi^2]\,(1-\lambda)}{[\lambda+\phi(1-\lambda)]^{\frac{\mu}{1-\sigma}}}\\&-\frac{2\phi\,(1-\lambda)+[1-\mu/\sigma+(1+\mu/\sigma)\phi^2]\,\lambda}{[(1-\lambda)+\phi\,\lambda]^{\frac{\mu}{1-\sigma}}}.\end{aligned}\tag{33}$$

Clearly, for the determination of equilibria all that matters is $\Delta V(\lambda,\phi)$. In particular, all interior equilibria are solutions to $\Delta V(\lambda,\phi)=0$ while fully agglomerated configurations $\lambda=0$ and $\lambda=1$ are equilibria if and only if $\Delta V(0,\phi)<0$ and $\Delta V(1,\phi)>0$ respectively. Since by (33) we have:

$$\Delta V(0,\phi)=-\Delta V(1,\phi)=\frac{[1-\mu/\sigma+(1+\mu/\sigma)\phi^2]}{\phi^{\frac{\mu}{1-\sigma}}}-2\phi$$

full agglomeration in either region is a stable spatial equilibrium whenever transport costs are so small that ϕ is above the threshold value ϕ_s, called the *sustain point*, which is implicitly defined by:

$$1-\mu/\sigma+(1+\mu/\sigma)\phi_s^2-2(\phi_s)^{1+\frac{\mu}{1-\sigma}}=0.\tag{34}$$

Turning to interior equilibria, we can prove that $\Delta V(\lambda,\phi)=0$ has at most three solutions for $0<\lambda<1$. It is readily verified that the symmetric outcome $\lambda=1/2$, which entails an even geographical distribution of skilled workers and firms, exists for any values of parameters. This solution is stable whenever $\Delta V_\lambda(1/2,\phi)<0$, where the subscript denotes the partial derivative with respect to the corresponding argument. This is the case

if and only if transport costs are sufficiently large for ϕ to be below the threshold value ϕ_b, called the *break point*, defined as:

$$\phi_b \equiv \frac{1 - 1/\sigma - \mu/\sigma}{1 - 1/\sigma + \mu/\sigma}\phi_r.$$

It is seen by inspection that the break point is decreasing in μ and increasing in σ. Moreover, if $\phi_b < 0$ the symmetric outcome is never stable and the market crowding effect is always dominated by market size and cost-of-living effects. We rule out this case by assuming that $\mu < \sigma - 1$ (the no-black-hole condition). Note also that the cost-of-living effect always works in favor of the large region. Therefore, at the break point, where real wages are equal, the smaller region must provide a wage premium.

Apart from $\lambda = 1/2$, there exist at most two other interior equilibria that are symmetrically placed around it. This comes from a tedious but standard study of the function $\Delta v(\lambda)$, which is symmetric around $\lambda = 1/2$ and changes concavity at most twice. In particular, the following local properties can be established in a neighborhood of $\lambda = 1/2$:[29]

$$\Delta v_\lambda(1/2, \phi_b) = 0 \qquad \Delta v_{\lambda\phi}(1/2, \phi_b) > 0 \tag{35}$$

$$\Delta v_{\lambda\lambda}(1/2, \phi_b) = 0 \qquad \Delta v_{\lambda\lambda\lambda}(1/2, \phi_b) > 0. \tag{36}$$

Because $\lambda = 1/2$ is always an equilibrium, $\Delta V(\lambda, \phi)$ rotates around $(\lambda, \Delta V) = (1/2, 0)$ as ϕ changes. Hence, (35) says that the steady state $\lambda = 1/2$ turns from stable to unstable as soon as ϕ grows above ϕ_b. Likewise, (36) says that, when the equilibrium $\lambda = 1/2$ changes stability, two additional equilibria emerge. Due to the symmetry of the model these equilibria are symmetric. All these properties together say that the differential equation (26) undergoes a (local) bifurcation at $\phi = \phi_b$. Moreover, the global extension of $\Delta v_{\lambda\lambda\lambda}(1/2, \phi_b) > 0$ implies that $\phi_b > \phi_s$.

The fact that transport costs at the break point are lower than at the sustain point implies that the model displays 'hysteresis' in location: once the CP equilibrium is reached, transport costs may rise above the break point before agglomeration ceases to be an equilibrium.

4.1.2 Quadratic utility and linear transport costs

The fact that (33) involves noninteger power variables makes the CP model not amenable to an analytical solution. By contrast, Ottaviano *et al.* (2002), who use the quasi-linear quadratic form (18) together with linear transport costs, have been able to obtain such a solution.[30] The indirect utility of a skilled worker is now given by:

$$v_i = S_i + w_i + \overline{q}_0 \qquad i = A, B$$

[29] Using the no-black-hole condition, all signs can be established by inspection.

[30] See Ludema and Wooton (2000) for a model of economic geography with $\beta = \gamma$ and quantity-setting oligopolistic firms.

where

$$S_i = \frac{a^2 N}{2b} - a\int_0^N p_i(s)ds + \frac{b+cN}{2}\int_0^N [p_i(s)]^2 ds - \frac{c}{2}\left[\int_0^N p_i(s)ds\right]^2$$

is the consumer surplus and w_i is the skilled wage.

The skilled wage prevailing in region A can be obtained by evaluating (20) at equilibrium prices, while taking into account that the numbers of consumers in the two regions are as follows:

$$M_A = \frac{L}{2} + \lambda H \qquad M_B = \frac{L}{2} + (1-\lambda)H$$

which yields the following expression:

$$w_A^*(\lambda, t) = \frac{b+cN}{f}\left[(p_{AA}-m)^2\left(\frac{L}{2}+\lambda H\right) + (p_{AB}-m-t)^2\left(\frac{L}{2}+(1-\lambda)H\right)\right] \tag{37}$$

with a symmetric expression holding in region B.

Using the equilibrium prices (21)-(22) as well as (37), the indirect utility differential is then:

$$\begin{aligned}\Delta v(\lambda, t) &\equiv S_A(\lambda, t) - S_B(\lambda, t) + w_A(\lambda, t) - w_B(\lambda, t) \\ &= D\, t(t^* - t)\cdot(\lambda - 1/2)\end{aligned} \tag{38}$$

where $D > 0$ is a bundle of parameters independent of t and

$$t^* \equiv \frac{4f(a-bm)(3bf+2cH)}{2bf(3bf+3cH+cL)+c^2H(H+L)} > 0.$$

It follows immediately from (38) that $\lambda = 1/2$ is always an equilibrium. Since $D > 0$, for $\lambda \neq 1/2$ the indirect utility differential has always the same sign as $\lambda - 1/2$ if and only if $t < t^*$; otherwise it has the opposite sign. When $t < t^*$, the symmetric equilibrium is unstable and workers agglomerate in region A (B) provided that the initial fraction of workers residing in this region exceeds $1/2$. In other words, agglomeration arises when transport costs are low enough, as in the foregoing and for similar reasons. In contrast, for large transport costs, that is, when $t > t^*$, it is straightforward to see that the symmetric configuration is the only stable equilibrium. Hence, the threshold t^* corresponds to both the critical value of t at which symmetry ceases to be stable (break point) and the value below which agglomeration is stable (sustain point); this follows from the fact that (38) is linear in λ.

When increasing returns are stronger, as expressed by higher values of f, t^* rises since $dt^*/df > 0$. This means that *the agglomeration of the manufacturing sector is more likely, the stronger are the increasing returns at the firm's level.* In addition, t^* increases with product differentiation since $dt^*/d\gamma < 0$. In words, *more product differentiation fosters agglomeration.*

It is readily verified that t^* is lower than t_{trade} when the population of unskilled is large relative to the population of skilled. Although the size of

the industrial sector is captured here through the relative population size of L/H and not through its share in consumption, the intuition is similar: the ratio L/H must be sufficiently large for the economy to display different types of equilibria according to the value of t, otherwise the coefficient of $(\lambda - 1/2)$ in (38) is always positive and agglomeration always prevails.[31] Our condition does not depend on the expenditure share on the manufacturing sector because of the absence of general equilibrium income effects: small or large sectors in terms of expenditure share may either be agglomerated when t is small enough.

The quadratic set-up allows us to derive analytically the results obtained by Krugman (1991). Nonetheless it should not be viewed as substitute but rather as complement to the CES set-up. While both models are not general, each has its own comparative advantage and should be used accordingly.[32]

The results of section 4.1 may be summarized as follows.

Proposition 4 (Core-periphery with Labor Migration) *Consider an economy with two regions and two sectors. The traditional sector employs only geographically immobile unskilled workers who are evenly distributed between regions. The modern sector employs also mobile skilled workers. Then, for low enough transport costs the only stable equilibrium has the whole modern sector agglomerated within the same region.*

In other words, for sufficiently low transport costs, even a small transitory shock to initially symmetric regions can give rise to large permanent regional imbalances. The different result with respect to Proposition 3 is due to the fact that some workers are now allowed to move whereas they all stay put in the setting considered in section 3. Individuals' mobility makes market sizes endogenous, thus allowing for the emergence of cumulative causation. When transport costs are low, the attraction of mobile consumers towards the larger market makes it even larger: as consumers relocate, its market access advantage grows whereas its market crowding disadvantage falls. Eventually, this leads to the agglomeration of all firms in one region. By contrast, when transport costs are high, the opposite holds in that market crowding now dominates market access, thus fostering the dispersion of firms.[33]

Such result concurs with Proposition 2 in which either sufficiently low transport costs, or a sufficiently high degree of product differentiation, or both are needed for a cluster of firms to arise in an oligopoly involving dispersed consumers. The fact that agglomeration arises when transport costs

[31] When varieties are complements, in which case $\gamma < 0$ and $c < 0$, it is readily verified that τ_{trade} is lower than τ^*, thus implying that agglomeration is the sole equilibrium.

[32] The CES functional form seems to be preferable whenever the total number of active firms depends on the their spatial distribution, as in the presence of vertical linkages (see 4.2). Conversely, the quasi-linear quadratic form is more desirable in symmetric settings when total factor endowments are given and the total number of firms is independent from their location (see section 5).

[33] When firms earn positive profits because entry is restricted, for some firms' ownership structures dispersion may become unstable for all transportation cost values. When this is the case, partial agglomeration arises for high transportation costs (Picard *et al.*, 2002).

are low also confirms what has been shown in spatial oligopoly theory (Irmen and Thisse, 1998). Therefore, it is tempting to answer Neary (2001, p. 551) that models of monopolistic competition seem to provide a reasonable approximation of what could be obtained in the (still missing) general equilibrium model with strategic interactions.[34]

4.2 The vertical linkage framework

In the foregoing, agglomeration arose because of the endogeneity of local market sizes due to mobile consumers. When labor is immobile across regions but perfectly mobile between sectors, the cumulative causation falls short and the symmetric equilibrium is the only stable outcome (Puga, 1999). Another reason for the market size to be endogenous is *the presence of input-output linkages between firms*: what is output for a firm is input for another and vice versa (the 'ancillarity' industries). In this case, the entry of a new firm in a region not only increases the intensity of competition between similar firms (market crowding effect); it also increases the size of the market of upstream firms-suppliers (market size effect) and decreases the costs of downstream firms-customers (*cost effect*).

The easiest way to introduce the above considerations is to model input-output linkages within the same industry.[35] Specifically, consider the model of section 4.1.1, but with two fundamental modifications.[36] First, there is only one factor of production, labor say, which is constant and in equal supply in the two regions:

$$L_i^Z + L_i^X = L/2$$

where $L_i^Z$and L_i^X are region i employments in sectors Z and X respectively. Labor can freely relocate between sectors within the same region but it is spatially immobile. As in the foregoing, this factor is used in both sectors to fulfill the variable input requirement. Again, when trade in the homogeneous good is costless, we have $w_i = p_i^Z = 1$ as long as each region is not

[34]Throughout this section, workers care only about their current utility level when choosing a location. This is a very restrictive assumption because migration decisions are typically made on the grounds of current *and* future utility flows as well as of various costs due to search, mismatch and homesickness. It is, therefore, important to figure out how the interplay between history and expectations shapes the space-economy when workers maximize the intertemporal value of their utility flows. Somewhat different approaches have been proposed to tackle this problem, but they yield similar conclusions (Ottaviano, 1999, 2001b; Baldwin, 2001; Ottaviano *et al.*, 2002): when the initial distribution of economic activities is not too skewed and when transport costs take intermediate values, the common belief that the skilled will eventually agglomerate in the smaller region can reverse the historically inherited advantage of the larger region. Hence, psychological forces may overcome the historical advantage of a region.

[35]In the original version of this model, Venables (1996) has an upstream and a downstream sectors. For simplicity, Krugman and Venables (1995) choose to collapse the two sectors into a single one. Here we present an analytically solvable version of Krugman and Venables (1995) due to Ottaviano (2002).

[36]See Ottaviano and Thisse (2001) for a treatment of the linear model with vertical linkages.

specialized in sector X, a condition that we assume to hold throughout this section. Second, *the fixed cost of manufacturing are incurred in a composite input consisting of labor and the differentiated varieties of good* X. For simplicity, as in Krugman and Venables (1995), the composite input is assumed to be Cobb-Douglas in L_i^X and Q_i with shares $1-\mu$ and μ respectively. Accordingly, the total cost function for a typical manufacturing firm is now given by - contrast with (2):

$$TC_i(s) = P_i^{\mu} w_i^{1-\mu} f + w_i m x_i(s)$$

where P_i is given by (11).

A typical firm located in region i maximizes profit:

$$\Pi_i(s) = p_{ii}(s)q_{ii}(s) + p_{ij}(s)q_{ij}(s) - m[q_{ii}(s) + tq_{ij}(s)] - P_i^{\mu} f \tag{39}$$

where we have again set $w_i = 1$. As a result, optimal pricing is still given by (10), which allows us to rewrite (39) as:

$$\Pi_i = \frac{m}{\sigma - 1} x_i - P_i^{\mu} f. \tag{40}$$

Intermediate demand implies that expenditures on manufactures now stem not only from consumers but also from firms:

$$\mu E_i = \mu Y_i + \mu n_i P_i^{\mu} f \tag{41}$$

where Y_i is consumers' income inclusive of firms' profits Π_i:

$$Y_i = \frac{L}{2} + n_i \Pi_i = \frac{L}{2} + n_i \left(\frac{m}{\sigma - 1} x_i - P_i^{\mu} f \right) \tag{42}$$

where the second equality is granted by (40). Then

$$\mu E_i = \mu \left(\frac{L}{2} + \frac{m}{\sigma - 1} n_i x_i \right).$$

Recalling (13), the X-sector market clearing condition becomes:

$$x_i = \frac{\sigma - 1}{m\sigma} \left(\frac{\mu E_i}{n_i + \phi n_j} + \frac{\phi \mu E_j}{\phi n_i + n_j} \right) \tag{43}$$

which, by (41) and (42), can be rewritten as follows:

$$x_i = \frac{\sigma - 1}{m\sigma} \left\{ \frac{\mu[L/2 + n_i x_i m/(\sigma - 1)]}{n_i + \phi n_j} + \frac{\phi \mu[L/2 + n_j x_j m/(\sigma - 1)]}{\phi n_i + n_j} \right\} \tag{44}$$

For $i = A, B$, (44) generates a system of linear equations that can be solved to obtain x_A and x_B as explicit functions of the numbers of active firms n_A and n_B. Standard derivations yield:

$$x_i = \frac{\sigma - 1}{m} \frac{\mu}{\sigma - \mu} \frac{L}{2} \frac{2\phi\, n_i + [1 - \mu/\sigma + (1 + \mu/\sigma)\phi^2]\, n_j}{\phi(n_i^2 + n_j^2) + [1 - \mu/\sigma + (1 + \mu/\sigma)\phi^2]\, n_i n_j}. \tag{45}$$

Thus, operating profits are equal to $w_i = (p_i - m)x_i$ as shown in (30).

We are now ready to analyze the entry decision of firms in the two regions. As before, we assume that agents are short sighted: firms enter when current profits are positive and exit when they are negative. Specifically, their flow is regulated by the following simple adjustment:

$$\frac{dn_i}{dt} = \begin{cases} \Pi_i(n_i, n_j, \phi) & \text{if } n_i > 0 \\ \max\{0, \Pi_i(n_i, n_j, \phi)\} & \text{if } n_i = 0 \end{cases} \tag{46}$$

where by (40) and (45):

$$\begin{aligned} \Pi_i(n_i, n_j, \phi) &= \frac{\mu/\sigma}{1-\mu/\sigma}\frac{L}{2}\frac{2\phi\, n_i + [1-\mu/\sigma+(1+\mu/\sigma)\phi^2]\, n_j}{\phi(n_i^2+n_j^2) + [1-\mu/\sigma+(1+\mu/\sigma)\phi^2]\, n_i n_j} \\ &\quad - \left(\frac{m\sigma}{\sigma-1}\right)^{\mu} (n_i + \phi n_j)^{\frac{\mu}{1-\sigma}} f. \end{aligned}$$

Unlike the CP model, *the vertical linkage model cannot be reduced to a unique differential equation* because the total number of firms is variable.

Given (46), a spatial equilibrium arises at (n_A^*, n_B^*), with $n_i^* > 0$, when $\Pi_i(n_i^*, n_j^*, \phi) = 0$ for $i = A, B$. It may also arise at $(n_A^*, n_B^*) = (n_A^o, 0)$, with $n_A^o > 0$, when $\Pi_A(n_A^o, 0, \phi) = 0$ and $\Pi_B(n_A^o, 0, \phi) < 0$; similar conditions define agglomeration in B. As in the CP model, equilibria may be multiple and stability is used to dismiss some of them.

Consider first an agglomerated configuration with all active firms in, say, region A: $n_A = n_A^o$ and $n_B = 0$. This is a stable equilibrium for (46) if and only if:

$$\Pi_A(n_A^o, 0, \phi) = 0 \tag{47}$$

so that no firm in A is willing to enter or exit, and

$$\Pi_B(n_A^o, 0, \phi) < 0 \tag{48}$$

so that no firm is willing to start production in B. It is readily verified that (47) is satisfied if and only if n_A^o is such that

$$n_A^o = \left[\frac{\mu/\sigma}{1-\mu/\sigma}\frac{L}{f}\left(\frac{\sigma-1}{m\sigma}\right)^{\mu}\right]^{\frac{1}{1+\frac{\mu}{1-\sigma}}}. \tag{49}$$

Moreover, (48) is met if and only if:

$$\frac{[1-\mu/\sigma+(1+\mu/\sigma)\phi^2]}{2\phi^{1+\frac{\mu}{1-\sigma}}} - 1 < 0$$

which implies that the condition for agglomeration to be a stable equilibrium in the present model is the same as in the CP model: ϕ must be larger than the sustain point ϕ_s as defined by (34).

Turning to interior equilibria, standard but tedious analysis reveals that the loci $\Pi_A(n_A, n_B, \phi) = 0$ and $\Pi_B(n_A, n_B, \phi) = 0$ always cross at least once

and no more than thrice. In particular, they always cross at the symmetric outcome in which n_A and n_B are both equal to:

$$n^* = 2^{-\frac{1-\sigma}{1-\sigma+\mu}}(1+\phi)^{-\frac{\mu}{1-\sigma+\mu}} n_A^0 \tag{50}$$

where n_A^o is given by (49) and independent of ϕ. As a result, while the number of active firms with agglomeration is invariant to trade barriers, the number of those active in the symmetric equilibrium is not.

The symmetric equilibrium is stable as long as the two eigenvalues of the Jacobian of (46) are negative at $n_A = n_B = n^*$. Taking into account the fact that the total output $n_A x_A + n_B x_B = [(\sigma-1)\mu L]/[(\sigma-\mu)m]$ is independent of the numbers of firms n_i gives a first eigenvalue $\partial\Pi_A/\partial n_A + \partial\Pi_B/\partial n_A$, which is always negative. As to the second eigenvalue $\partial\Pi_A/\partial n_A - \partial\Pi_B/\partial n_A$, its sign is negative (resp., positive) whenever $\phi <$ (resp., $>$) ϕ_b. Thus, in the vertical-linkage model we have not only the same sustain point (ϕ_s) and the same break point (ϕ_b) but also the same no-black-hole condition ($\mu < \sigma - 1$) as in the CP model.

Finally, under the no-black-hole condition, (50) reveals that n^* is an increasing function of ϕ. The freer trade is, the larger the number of active firms: trade integration fragments the market. Moreover, under symmetry the number of active firms in each region is larger than the total number of firms under agglomeration ($n^* > n_A^o$). Therefore, differently from the CP framework, in the vertical-linkage model, *agglomeration defragments the market and reduces product variety.*

To sum up, also with vertical linkages, small transitory shocks can have large permanent effects:

Proposition 5 (Core-periphery with Vertical Linkages) *Consider an economy with two regions and two sectors. The traditional sector employs only geographically immobile workers who are evenly distributed between regions. The modern sector employs also intermediate inputs supplied by the industrial firms. Then, for low enough transport costs the only stable equilibrium has the whole modern sector agglomerated within the same region.*

This result differs from Krugman and Venables (1995) who obtain the following pattern as transport costs fall: dispersion-agglomeration-redispersion. This difference is due to the fact that we assume here that regions never fully specialize in the production of the differentiated good. This is not an innocuous assumption because it insures that the agglomeration of firms within a region does not trigger wage divergence between regions. *When wages increase with the number of firms setting in the core, beyond some threshold freer trade may lead to the redispersion of the modern sector because local wages are too high.* Giving a full analytical treatment of this model is a task beyond our reach and we refer the reader to Puga (1999) for what is probably the best analysis of this model and of several other generalizations.

5 The bell-shaped curve of spatial development

The NEG models surveyed in sections 3 and 4 provide possible explanations for a certain number of stylized facts summarized in the introduction. Yet, they rest on a set of very peculiar assumptions. Moreover, the explanation of other stylized facts remains beyond their reach. In this section we consider the most important of the unexplained facts to show that NEG models can be made both less theoretically restrictive and more empirically appealing by removing some of their most peculiar assumptions.

In terms of stylized facts, there exists a strand of literature that connects the evolution of the spatial distribution of population and industry to the various stages of economic development (Williamson, 1965; Wheaton and Shishido, 1981). These authors argue that a high degree of urban concentration together with a widening urban-rural wage differential is expected to arise during the early phases of economic growth. As development proceeds, spatial deconcentration and a narrowing wage differential should occur. Hence, *the emergence of a core-periphery structure would be followed by a phase involving interregional convergence.*[37]

Such a bell-shaped relation between the degree of spatial concentration of economic activities and the degree of goods and factors mobility (the so-called 'bell-shaped curve of spatial development') does not arise in either the HME nor the CP models presented so far. Here we survey some recent contributions that overcome that limitation. For ease of exposition, we name the traditional (resp., modern) sector 'agriculture' (resp., 'industry') as in Krugman (1991).

5.1 More on spatial costs

A first somewhat awkward assumption of the models of sections 3 and 4 is that the transport cost of one good is taken into account whereas the transport cost of the other is neglected. There is no evidence that shipping a differentiated good costs much more than shipping a homogeneous good (Rauch, 1996). The role of the traditional sector is very modest in the HME and CP models: it permits trade imbalances in the industrial good. A second unsatisfactory assumption is that the agglomeration of workers into a single region does not involve any agglomeration costs. Yet, it is reasonable to believe that a growing settlement in a given region will often take the form of an urban area, typically a city in which land becomes a critical commodity. In what follows we remove the assumptions sequentially and show that in both scenarios a bell-shaped curve of spatial development may indeed emerge. Interestingly, in each case, what is crucial for the spatial organization of the economy is *the value of the transport cost relative to the additional spatial cost taken into consideration.*

[37]Note that the evidence discussed in the chapter by Kim and Margo in this volume seems to support the idea that a bell-shaped curve fits the case of the United States. Although some other developed economies would also seem to experience re-dispersion (Geyer and Kontuly, 1996), more empirical work on the issue would be certainly welcome.

Surprisingly enough, dealing with transport costs in the case of two sectors does not appear to be a simple task. In the model proposed by Krugman (1980), which is very similar to the one considered in section 3.2.1, Davis (1998) shows that *the HME vanishes when both the industrial and the agricultural goods are shipped at the same cost.* The intuition is fairly straightforward. Suppose that both regions produce their own requirements of the agricultural good, so that industry is distributed according to region size. If some firms are relocated into the large region, then trade in the industrial good falls whereas trade in the agricultural good rises. Transport costs being the same for the two goods, one expects total transport costs in the economy to rise, thus implying that the shifted firms find the move unprofitable.

The issue is more involved in CP models. To our knowledge, the first analytical solution has been provided by Picard and Zeng (2003) who use the linear model of section 4.1.2. For simplicity, assume that the agricultural good is differentiated and that region $i = A, B$ is specialized in the production of variety i.[38] Preferences are quadratic and given by

$$U_i^a = U_i + \alpha_a(q_A^a + q_B^a) - \frac{\beta_a - \gamma_a}{2}[(q_A^a)^2 + (q_B^a)^2] - \frac{\gamma_a}{2}(q_A^a + q_B^a)^2$$

where U_i is given by (18) and q_i^a the quantity of variety i of the agricultural good. As usual, the first order conditions yields the demands for this good in region A as follows:

$$q_{ji}^a = a_a - (b_a + 2c_a)\, p_{ji}^a + c_a(p_{ii}^a + p_{ji}^a)$$

where

$$a_a \equiv \frac{\alpha_a}{\beta_a + \gamma_a} \qquad b_a \equiv \frac{1}{\beta_a + \gamma_a} \qquad c_a \equiv \frac{\rho_a}{(\beta_a - \gamma_a)(\beta_a + \gamma_a)}$$

Shipping one unit of each variety of the agricultural good requires $t_a > 0$ units of the numéraire. Picard and Zeng (2003) then show that the utility differential (38) becomes

$$\Delta v(\lambda, t, t_a) = U_A^a - U_B^a = \left[D\, t(t^* - t) + G t_a^2\right] \cdot (\lambda - 1/2)$$

where $G > 0$ is a bundle of parameters independent of t and t_a. For any given t_a, the expression $D\, t(t^* - t) + G t_a^2$ may have one, two or no zeros. Clearly, there exists a value of t_a for which it has a single zero, which is denoted $\overline{t}_a$. When $t_a < \overline{t}_a$, the equation $Dt(t^* - t) + G t_a^2 = 0$ has two real and distinct roots $\overline{t}_1$ and $\overline{t}_2$. Hence, it is immediate to show the following result:

Proposition 6 *(***Shipping the Agricultural Good Is Costly***) Consider the CP model with labor migration of section 4.1. If agricultural shipping*

[38] When the agricultural good is homogenous, dispersion always prevails as soon as τ_a is positive (Picard and Zeng, 2003). See also Fujita *et al.* (1999, ch.7).

costs are low, then the market equilibrium has all industrial firms located in the same region whenever their own transport costs take intermediate values; otherwise, the symmetric configuration is the only stable spatial equilibrium. By contrast, if agricultural shipping costs are large enough, dispersion always prevails.

In the present context, economic integration has two dimensions described by the parameters t and t_a. When shipping the agricultural good is sufficiently cheap, the spatial distribution of industry is bell-shaped with respect to its own transport cost but changes in regimes remain catastrophic. It is interesting to point out that while dispersion arises for both high and low transport costs, this happens for very different reasons. In the former case, firms are dispersed due to the crowding effect on the industrial good market; in the latter, the working force is the price differential of the agricultural good. A new result also emerges: when shipping the agricultural good is expensive, industry is always dispersed. All of these agree with the simulations reported on by Fujita *et al.* (1999, ch. 7) for a similar extension of Krugman's (1991) model and show that *the level of the agricultural good's transport costs matters for the location of industrial firms.*[39]

5.2 Heterogenous workers

A third unappealing assumption of the NEG models considered in sections 3 and 4 is that individuals have the same preferences. Although this assumption is not uncommon in economic modeling, it seems highly implausible that all potentially mobile individuals will react in the same way to a given real wage gap between regions. Some people show a high degree of attachment to the region where they were born; they will stay put even though they may guarantee to themselves higher real wage in other places. In the same spirit, life-time considerations such as marriage, divorce and the like play an important role in the decision to migrate. Finally, regions are not identical and exhibit different natural and cultural features. Typically, individuals differ in their reactions to these various factors. Although individual motivations are difficult to model, Tabuchi (1986) has argued that it is possible to identify their aggregate impact on the spatial distribution

[39]Tabuchi (1998) and Ottaviano *et al.* (2002) assume that the (main) dispersion force rests on land consumption, which rises with the size of the population established within the same region whereas firms set up in a regional employment center. Using the two models of section 4.1, these authors show that the existence of commuting costs within each regional cluster is sufficient to yield dispersion when the transport costs are sufficiently low. Hence, as transport costs fall, the economy involves dispersion, agglomeration, and re-dispersion. In other words, sufficiently high commuting costs always yield dispersion. This is strikingly similar to Proposition 6, but re-dispersion is now triggered by the crowding of the land market. In the present context, we may assume that there is no farmers without annihilating the dispersion force. If $L = 0$, it is readily verified that the economy moves from agglomeration to dispersion when trade costs fall, thus confirming the numerical results obtained by Helpman (1998). Note also that a bell-shaped curve arises in the vertical linkage model of Krugman and Venables (1995) when unskilled labor is replaced by land.

of economic activities by using discrete choice theory. Stated differently, a discrete choice model can be used to capture the matching between individuals and regions. Tabuchi and Thisse (2002b) have then combined the CP model of 4.2.2 with the logit in order to asses the impact of heterogeneity in migration behavior.[40]

Let $v_r(\lambda)$ be the indirect utility obtained from consuming the industrial and agricultural goods in region r. Then, the probability that a worker will choose to reside in region r is given by the logit formula:

$$\mathbf{P}_r(\lambda) = \frac{\exp[v_r(\lambda)/\upsilon]}{\exp[v_r(\lambda)/\upsilon] + \exp[v_s(\lambda)/\upsilon]} \tag{51}$$

In (51), υ expresses the dispersion of individual tastes: the larger υ, the more heterogenous the workers' tastes about their living place. When $\upsilon = 0$, workers are homogenous and behave as in section 4.1. In the present setting, it should be clear that the population of skilled workers changes according to the following equation of motion:

$$\frac{d\lambda}{dt} = (1-\lambda)\mathbf{P}_A(\lambda) - \lambda\mathbf{P}_B(\lambda) \tag{52}$$

where the first term in the RHS of (52) stands for the fraction of people migrating into region A, while the second term represents those leaving this region for region B. A spatial equilibrium λ^* arises when $d\lambda/dt = 0$. Tabuchi and Thisse (2002b) show that the sign of $\partial\lambda^*/\partial t$ is identical to the sign of $Dt(t^* - t) - 4\upsilon$. It is then readily verified that $\upsilon = \upsilon^* \equiv D(t^*)^2/16$ is a zero of the discriminant of this expression. Accordingly, when $\upsilon < \upsilon^*$, the equation $Dt(t^* - t) - 4\upsilon = 0$ has two real and distinct real roots t_1^* and t_2^*. The following result then holds.

Proposition 7 *(***Probabilistic Migration Behavior***) Consider the CP model with labor migration of section 4.1. Assume that workers differ in terms of attachment to the two regions. If workers' heterogeneity is small, the market equilibrium involves full dispersion for high transport costs; when transport costs take intermediate values, the industry is partially agglomerated, with the gap between regions being a bell-shaped function of transport costs; finally, the industry is again fully dispersed once transport costs are sufficiently low. If workers heterogeneity is large, complete dispersion always prevails.*

In other words, when $t < t_1^*$ workers are dispersed as long as $\upsilon > 0$, whereas they would be agglomerated in the standard case because $t_1^* = 0$ and $t_2^* = t^*$ for $\upsilon = 0$. When heterogeneity is positive but weak, industry displays a *smooth bell-shaped* pattern. Furthermore, full agglomeration never arises and the economy moves away from dispersion in a noncatastrophic manner. By contrast, when υ is large enough, there is always dispersion. Hence,

[40] This assumption turns out to be empirically relevant in migration modeling (Anderson and Papageorgiou, 1994), while it is analytically convenient without affecting the qualitative nature of the main results. See also section 2.2.3.

we may then conclude that taste heterogeneity is a strong dispersion force that deeply affects the formation of the space-economy.[41] Thus, workers' heterogeneity has the same type of impact on the space-economy as positive commuting costs within each region and positive transport costs for the agricultural good. The main difference lies in the smoothness of the process.

6 Where do we go from here

NEG models investigate the nature of the interplay between imperfect competition, plant-level returns to scale and the associated pecuniary externalities. They show that even small transitory shocks can have large permanent effects on the economic landscape. This is consistent with the emergence of a *putty-clay economic geography* that seems to be one of the main features of modern economies: the steady fall in transport costs seems to allow for a great deal of flexibility on where particular activities can locate, but once spatial differences develop, locations tend to become quite rigid. Hence, even in the absence of major events, *regions that were once similar may end up having very different production structures.* Nonetheless, we have also seen that such an extreme agglomeration may give rise to various forms of price differentials that can trigger a process of redispersion, or that more sophisticated migration behavior may prevent the emergence of a single core. Thus, it is fair to say that no definitive conclusion emerges, even though the case for agglomeration seems to be strong.

Several issues remain untouched or are, at the very least, poorly understood. First, more attention should be paid to the nature and type of the various forces at work in economic geography. For example, rather unexpectedly, some recent contributions have shown that using a population of immobile workers (as in section 4) or urban costs growing with the population size as a dispersion force, may lead to fairly contrasted results (Krugman and Livas Elizondo, 1996; Helpman, 1998; Monfort and Nicolini, 2000; Tabuchi and Thisse, 2002b). This invites us to investigate more precisely *how, but also why, different agglomeration and dispersion forces may affect the distribution of economic activities.* The importance of the forces at stake may vary with time and place, but also with the stage of development as well as with the spatial scale.

Someway paradoxically, NEG models have fallen short of full-fledged welfare analysis.[42] One exception can be found in Ottaviano and Thisse (2002), who provide a new welfare analysis of agglomeration. They argue that, while natural due to the many market imperfections that are present in new economic geography models, such an analysis has been seldom touched due to the limits of the standard CES approach. What they show is that the market yields agglomeration for values of the transport costs for which it is socially desirable to keep activities dispersed. In particular, while they

[41] Similar conclusions are obtained by Murata (2003) for the Dixit-Stiglitz-iceberg model.

[42] See, however, Krugman and Venables (1995) and Helpman (1998) for some numerical developments about the welfare implications of agglomeration in related models.

coincide for high and low values of the transport costs, *the equilibrium and the optimum differ for a domain of intermediate values.* This opens the door to regional policy interventions grounded on both efficiency and equity considerations. More work is called for here to confirm this preliminary result.[43]

Another drawback of the NEG models is the assumption of a two-region setting borrowed from trade theory. By their very nature, such models are unable to explain the rich and complex hierarchy that characterizes the space-economy. In addition, if they allow one to better understand *why* agglomeration occurs, those models have little to say about *where* it arises. Therefore, one major step on the research agenda is the study of a multiregional system whose aim is to understand why some regions are more successful than others. To accomplish this task, we need to account for the actual geography of these regions, something that trade theorists have put aside. Moving into this direction also requires the study of the role of non-market institutions, such as organized interests and the polity, in the process of spatial development. This is a hard but exciting task.[44]

Another fundamental question is related to the fact that local labor markets are modeled in a very simple way in NEG: operating profits are used to pay the skilled workers.[45] In particular, these models do not help understand why unemployment persists in areas included in or adjacent to prosperous regions. Given the social importance of the issue at stake, one should investigate the possible reasons for the existence of a geography of employment that shows strong regional disparities. In addition, it is worth stressing that labor markets are probably the most regulated ones in Europe, with national institutions varying sometimes substantially across countries. Here too, we ignore what the impacts of such institutions are on the spatial organization of economic activities. Last, labor mobility between sectors should be studied in relation to the way regional prospects interfere with the training level of local workers. Clearly, more work is called for in these various directions for the NEG models to permit the design of what could be more effective regional development policies.[46]

Finally, NEG has very much relied on the development of new, but specific, models of monopolistic competition. Future progress in the field (and in others) will largely depend on the economists' ability to tame increasing

[43] Other recent contributions to the welfare analysis of agglomeration can be found in Baldwin *et al.* (2003) and Charlot *et al.* (2003).

[44] A first attempt in this direction is Robert-Nicoud and Sbergami (2003) who show how a political process may weaken, or even reverse, the home-market effect studied in section 3.

[45] Picard and Toulemonde (2002) offer an alternative, promising way to deal with some aspects of the labor market.

[46] Baldwin *et al.* (2003) provide a detailed survey of the various policy implications of NEG models grounded on welfare analysis. In particular, they show that many standard regional policies, which often use *ceteris paribus* arguments, may well deliver outcomes that vastly differ from those expected. By taking into account several general equilibrium effects usually neglected, these models provide a new framework to think about the design of regional policies.

returns and imperfect competition within more general models than those used so far. Building such models is a 'must' that will require much more imagination.

References

[1] Anderson S.P., A. de Palma and J.-F. Thisse (1992) Discrete choice theory of product differentiation (MIT Press, Cambridge, MA).

[2] Anderson W.P. and Y.Y. Papageorgiou (1994) An analysis of migration streams for the Canadian regional system, 1952-1983. 1. Migration probabilities, Geographical Analysis 26, 15-36.

[3] Baldwin R.E. (2001) Core-periphery model with forward-looking expectations, Regional Science and Urban Economics 31, 21-49.

[4] Baldwin R.E., R. Forslid, Ph. Martin, G.I.P. Ottaviano and F. Robert-Nicoud (2003) Economic geography and public policy (Princeton University Press, Princeton).

[5] Beckmann M.J. and J.-F. Thisse (1986) The location of production activities, in: P. Nijkamp, ed.,. Handbook of regional and urban economics, Vol. 1 (North-Holland, Amsterdam) 21-95.

[6] Behrens K. (2002) Agglomeration under autarky: when goods are immobile, LATEC, Université de Bourgogne, mimeo.

[7] Chamberlin E.H. (1933) The theory of monopolistic competition (Harvard University Press, Cambridge, MA).

[8] Charlot S., C. Gaigné, F. Robert-Nicoud, and J.-F. Thisse (2003) Agglomeration and Welfare: the core-periphery model in the light of Bentham, Hicks, Kaldor, and Rawls, INRA, mimeo.

[9] d'Aspremont C., J. J. Gabszewicz and J.-F. Thisse (1979) On Hotelling's "Stability in Competition", Econometrica 47, 1045-1050.

[10] Davis D.R. (1998) The home market effect, trade, and industrial structure, American Economic Review 88, 1264-1276.

[11] de Palma A., V. Ginsburgh, Y.Y. Papageorgiou and J.-F. Thisse (1985) The principle of minimum differentiation holds under sufficient heterogeneity, Econometrica 53, 767-781.

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

[13] Eaton B.C. and R.G. Lipsey (1977) The introduction of space into the neoclassical model of value theory, in: M. Artis and A. Nobay, eds., Studies in modern economics (Basil Blackwell, Oxford) 59-96.

[14] Faini R. (1999) Trade unions and regional development, European Economic Review 43, 457-474.

[15] Forslid R. and G.I.P. Ottaviano (2003) An analytically solvable core-periphery model, Journal of Economic Geography 3, forthcoming.

[16] Fujita M. (1988) A monopolistic competition model of spatial agglomeration: A differentiated product approach, Regional Science and Urban Economics 18, 87-124.

[17] Fujita M., P. Krugman and A.J. Venables (1999) The spatial economy. Cities, regions and international trade (MIT Press, Cambridge, MA).

[18] Fujita M. and J.-F. Thisse (2002) Economics of agglomeration. Cities, industrial location, and regional growth (Cambridge University Press, Cambridge).

[19] Gabszewicz J. J. and J.-F. Thisse (1986) Spatial competition and the location of firms, in: J.J. Gabszewicz, J.-F. Thisse, M. Fujita and U. Schweizer, Location theory (Harwood Academic Publishers, Chur) 1-71.

[20] Gabszewicz J. J. and J.-F. Thisse (1992) Location, in: R.E. Aumann and S. Hart, eds., Handbook of game theory with economic applications, Vol. 1 (North-Holland, Amsterdam) 281-304.

[21] Geyer H.S. and T.M. Kontuly (1996) Differential urbanization: integrating spatial models (Arnold, London).

[22] Ginsburgh V., Y.Y. Papageorgiou and J.-F. Thisse (1985) On existence and stability of spatial equilibria and steady-states, Regional Science and Urban Economics 15, 149-158.

[23] Greenhut M.L. (1981) Spatial pricing in the United States, West Germany and Japan, Economica 48, 79-86.

[24] Greenhut M.L., G. Norman and C.-S. Hung (1987) The economics of imperfect competition. A spatial approach (Cambridge University Press, Cambridge).

[25] Haig R.M. (1926) Toward an understanding of the metropolis. I. Some speculations regarding the economic basis of urban concentration, Quarterly Journal of Economics 40, 179-208.

[26] Hansen P., M. Labbé, D. Peeters and J.-F. Thisse (1987) Facility location analysis, in: P. Hansen, M. Labbé, D. Peeters, J.-F. Thisse, and J.V. Henderson, Systems of cities and facility location (Harwood Academic Publishers, Chur) 1-70.

[27] Head K. and T. Mayer (2000) Non-Europe. The magnitude and causes of market fragmentation in the EU, Weltwirtschaftliches Archiv 136, 284-314.

[28] Head K., T. Mayer, and J. Ries (2002) On the pervasiveness of the home market effect, Economica 69, 371-390.

[29] 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 University Press, Cambridge) 33-54.

[30] Helpman E. and P.R. Krugman (1985) Market structure and foreign trade, (MIT Press, Cambridge, MA).

[31] Henderson J.V. (1988) Urban development. Theory, fact and illusion (Oxford University Press, Oxford).

[32] Hotelling H. (1929) Stability in competition, Economic Journal 39, 41-57.

[33] Irmen A. and J.-F. Thisse (1998) Competition in multi-characteristics spaces: Hotelling was almost right, Journal of Economic Theory 78, 76-102.

[34] Isserman A.M. (1995) "It's Obvious, It's Wrong, and Anyway They Said It Years Ago"? Paul Krugman and large cities, International Regional Science Review 19, 37-48.

[35] Kaldor N. (1935) Market imperfection and excess capacity, Economica 2, 35-50.

[36] Koopmans T.C. (1957) Three essays on the state of economic science (McGraw-Hill, New York).

[37] Krugman P.R. (1980) Scale economies, product differentiation, and the pattern of trade, American Economic Review 70, 950-959.

[38] Krugman P.R. (1991) Increasing returns and economic geography, Journal of Political Economy 99, 483-499.

[39] Krugman P.R. (1993) The hub effect: or, threeness in international trade, in: W.J. Ethier, E. Helpman and J.P. Neary, eds., Theory, policy and dynamics in international trade (Cambridge University Press, Cambridge).

[40] Krugman P.R. and R. Livas Elizondo (1996) Trade policy and the third world metropolis, Journal of Development Economics 49, 137-150.

[41] Krugman P. R. and A.J. Venables (1995) Globalization and the inequality of nations, Quarterly Journal of Economics 60, 857-880.

[42] Kuehn A.A. and M.J. Hambuger (1963) A heuristic program for locating warehouses, Management Science 9, 643-666.

[43] Lampard E.E. (1955) The history of cities in the economically advanced areas, Economic Development and Cultural Change 3, 321-342.

[44] Lerner A. and H.W. Singer (1937) Some notes on duopoly and spatial competition, Journal of Political Economy 45, 145-186.

[45] Lösch A. (1940) Die Räumliche Ordnung der Wirtschaft (Gustav Fischer, Jena). English translation: The economics of location (Yale University Press, New Haven, CN, 1954).

[46] Ludema R.D. and I. Wooton (2000) Economic geography and the fiscal effects of regional integration, Journal of International Economics 52, 331-352.

[47] Martin Ph. and C.A. Rogers (1995) Industrial location and public infrastructure, Journal of International Economics 39, 335-351.

[48] Martin R. (1999) The new 'geographical turn' in economics: some critical reflections, Cambridge Journal of Economics 23, 65-91.

[49] Matsuyama K. (1995) Complementarities and cumulative process in models of monopolistic competition, Journal of Economic Literature 33, 701-729.

[50] McFadden D. (1974) Conditional logit analysis of qualitative choice behavior, in: P. Zarembka, ed., Frontiers in econometrics (Academic Press, New York) 105-142.

[51] Monfort Ph. and R. Nicolini (2000) Regional convergence and international integration, Journal of Urban Economics 48, 286-306.

[52] Murata Y. (2003) Product diversity, taste heterogeneity, and geographic distribution of economic activities: Market vs. non-market interactions, Journal of Urban Economics, forthcoming.

[53] Myrdal G. (1957) Economic theory and underdeveloped regions (Duckworth, London).

[54] Neary J.P. (2001) Of hype and hyperbolas: introducing the new economic geography, Journal of Economic Literature 39, 536-561.

[55] Ohlin B., P.O. Hesselborn and P.M. Wijkman, eds. (1977) The international allocation of economic activity (Macmillan, London).

[56] Ottaviano G.I.P. (1999) Integration, geography, and the burden of history, Regional Science and Urban Economics 29, 245-256.

[57] Ottaviano G.I.P. (2001a) Home market effects and the (in)efficiency of international specialization, GIIS, mimeo.

[58] Ottaviano G.I.P. (2001b) Monopolistic competition, trade, and endogenous spatial fluctuations, Regional Science and Urban Economics 31, 51-77.

[59] Ottaviano, G.I.P.(2002) Models of 'new economic geography': factor mobility vs. vertical linkages, GIIS, mimeo.

[60] Ottaviano G.I.P., T. Tabuchi and J.-F. Thisse (2002) Agglomeration and trade revisited, International Economic Review 43, 409-436.

[61] Ottaviano G.I.P. and J.-F. Thisse (2001) On economic geography in economic theory: increasing returns and pecuniary externalities, Journal of Economic Geography 1, 153-179.

[62] Ottaviano G. and J.-F. Thisse (2002) Integration, agglomeration and the political economics of factor mobility, Journal of Public Economics 83, 429-456.

[63] Papageorgiou Y.Y. and J.-F. Thisse (1985) Agglomeration as spatial interdependence between firms and households, Journal of Economic Theory 37, 19-31.

[64] Picard P. and E. Toulemonde (2002) Endogenous qualifications and firms' aggomeration, CORE Discussion Paper N°2002/70.

[65] Picard P. and E. Toulemonde (2003) Regional asymmetries: economies of agglomeration versus unionized labor markets, Regional Science and Urban Economics 33, 223-249.

[66] Picard P., E. Toulemonde and J.-F. Thisse (2002) Economic geography and the role of profits, CEPR Discussion Paper N°3385.

[67] Picard P. and D.-Z. Zeng (2003) The agricultural sector in the core-periphery model, CORE Discussion Paper N°.

[68] Pollard S. (1981) Peaceful conquest. The industrialization of Europe 1760-1970 (Oxford University Press, Oxford).

[69] Puga D. (1999) The rise and fall of regional inequalities, European Economic Review 43, 303-334.

[70] Rauch J.E. (1996) Network versus markets in international trade, NBER Discussion Paper N°5617.

[71] Reilly W.J. (1931) The law of retail gravitation, (Pilsbury, New York).

[72] Robert-Nicoud F. and F. Sbergami (2003) Home-market vs. vote-market effect: location equilibrium in a probabilistic voting model, European Economic Review, forthcoming.

[73] Sakashita N. (1967) Production function, demand function and location theory of the firm, Papers and Proceedings of the Regional Science Association 20, 109-129.

[74] SOPEMI (1998) Trends in international migration (OECD, Paris).

[75] Starrett D. (1978) Market allocations of location choice in a model with free mobility, Journal of Economic Theory 17, 21-37.

[76] Stahl K. (1983) A note on the microeconomics of migration, Journal of Urban Economics 14, 318-326.

[77] Stahl K. (1987) Theories of urban business location, in: E.S. Mills, ed., Handbook of regional and urban economics, Vo. 2 (North-Holland, Amsterdam) 759-820.

[78] Tabuchi T. (1986) Existence and stability of city-size distribution in the gravity and logit models, Environment and Planning A 18, 1375–1389.

[79] Tabuchi T. (1998) Agglomeration and dispersion: a synthesis of Alonso and Krugman, Journal of Urban Economics 44, 333-351.

[80] Tabuchi T. and J.-F. Thisse (2002a) Taste heterogeneity, labor mobility and economic geography, Journal of Development Economics 69, 155-177.

[81] Tabuchi T. and J.-F. Thisse (2002b) Regional specialization and transport costs, CEPR DP N°3542.

[82] Thisse J.-F. and X. Vives (1988) On the strategic choice of a spatial price policy, American Economic Review 78, 122-137.

[83] Thomas I. (2002) Transportation networks and the optimal location of human activities. A numerical geography approach (Edward Elgar, Cheltenham, UK).

[84] Tirado D.A., E. Paluzie and J. Pons (2002) Economic integration and industrial location: the case of Spain before World War I. Journal of Economic Geography 2, 343-363.

[85] Venables A.J. (1996) Equilibrium locations of vertically linked industries, International Economic Review 37, 341-359.

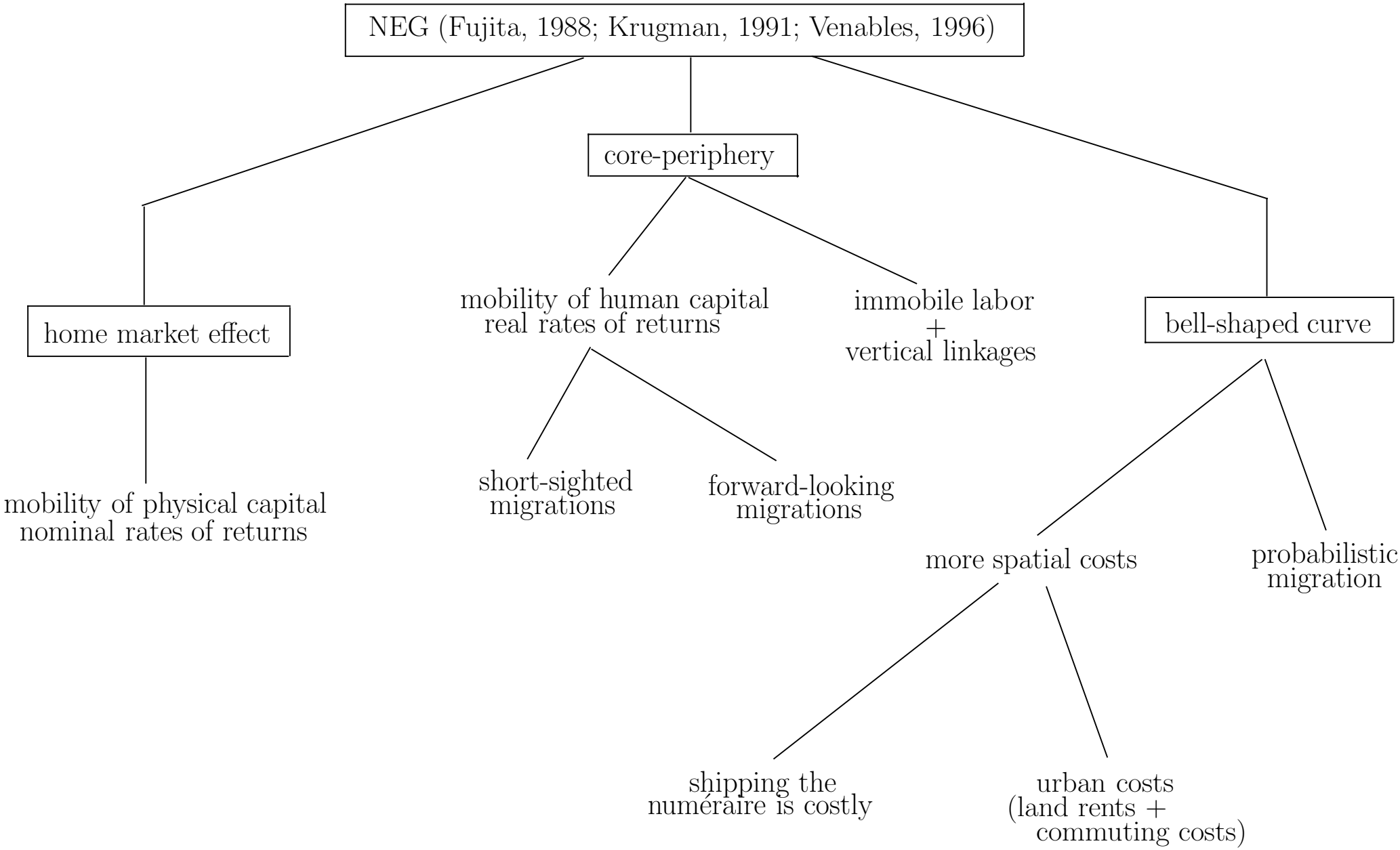
[86] Weber A. (1909) Über den Standort der Industrien (J.C.B. Mohr, Tübingen). English translation: The theory of the location of industries (Chicago University Press, Chicago, 1929).

[87] Weiszfeld E. (1936) Sur le point pour lequel la somme des distances de n points donnés est minimum, Tôhoku Journal of Mathematics 43, 355-386.

[88] Wheaton W. and H. Shishido (1981) Urban concentration, agglomeration economies, and the level of economic development, Economic Development and Cultural Change 30, 17-30.

[89] Williamson J. (1965) Regional inequality and the process of national development, Economic Development and Cultural Change 14, 3-45.

[90] Witzgall C. (1964) Optimal location of a central facility: mathematical models and concepts, National Bureau of Standard Report 8388, Washington D.C.


NEG (Fujita, 1988; Krugman, 1991; Venables, 1996)
core-periphery
home market effect
mobility of physical capital
nominal rates of returns
mobility of human capital
real rates of returns
immobile labor
+
vertical linkages
bell-shaped curve
short-sighted
migrations
forward-looking
migrations
more spatial costs
probabilistic
migration
shipping the
numéraire is costly
urban costs
(land rents +
commuting costs)