International Regional Science Review

http://irx.sagepub.com

The Role of Geography in Development

Paul Krugman

International Regional Science Review 1999; 22; 142
DOI: 10.1177/016001799761012307

INTERNATIONAL REGIONAL SCIENCE REVIEW 22, 2: 142–161 (August 1999)

THE ROLE OF GEOGRAPHY IN DEVELOPMENT

PAUL KRUGMAN
Department of Economics, Massachusetts Institute of Technology, Cambridge, MA, krugman@mit.edu

This article assesses how the tension between centripetal forces (such as forward and backward linkages in production and increasing returns in transportation) and centrifugal forces (such as factor immobility and land rents) can produce a process of self-organization in which symmetric locations end up playing very different economic roles. The article discusses geographic models of the division of the world into industrial and developing countries, of the emergence of regional inequality within developing countries, and of the emergence of giant urban centers. It argues that the conflict between "predestination" and "self-organizing" approaches to economic geography may be more apparent than real and briefly discusses policy—mainly in terms of why it is so hard to draw policy conclusions from economic geography models.

In recent years there has been a surge of interest in the geographic aspects of development, that is, in the question of where economic activities take place. There is nothing surprising about this interest—or perhaps the surprise is that it took so long for this interest to become a mainstream concern within economics. After all, even a casual look at a map of the world suggests that differences in economic development are at the very least associated with location: countries close to the equator tend to be poorer than those in temperate zones, and per capita income in Europe seems to follow a downward gradient from the northwest corner of the continent. It also is apparent that there are large regional inequalities within countries and, often, a powerful tendency for populations to concentrate in a few densely populated regions and cities. But only recently have attempts to explain such patterns become a subject for research by a large number of economists.

The new interest in economic geography usually takes one of two seemingly contradictory approaches. One approach—exemplified by John Luke Gallup, Jeffrey Sachs, and Andrew Mellinger's article in this issue—attempts to explain the differences in economic development between locations in terms of underlying, inherent differences in those locations. That is, it looks for associations such as the tendency of countries with tropical climates to have low per capita income or of great cities to emerge where there are good harbors.

TABLE 1. Forces Affecting Geographic Concentration

Centripetal Forces	*Centrifugal Forces*
Market size effects (linkages)	Immobile factors
Thick labor markets	Land rents
Pure external economies	Pure external diseconomies

The other approach typically asks why the economic destinies of locations might diverge even in the absence of such inherent advantages or disadvantages, why small historical accidents can cause one country to become part of the industrial core while another becomes part of the primary-producing periphery, or why some more or less arbitrary location becomes the site of a megacity containing ten million or more people. These two approaches may well seem contradictory: one seems to be a story of predestination, the other a story of chance. As I argue later in this article, however, the contradiction is more apparent than real. In fact, understanding why small random events can have large consequences for economic geography also is crucial to understanding why underlying differences in natural geography can have such large effects. Thus, the two approaches turn out to be complementary rather than contradictory.

In any case, most of this article is devoted to understanding how the geography of the world economy—both between and within nations—can engage in a process of self-organization in which locations with seemingly identical potential end up playing very different economic roles.

THEORETICAL PRINCIPLES OF THE NEW ECONOMIC GEOGRAPHY

Many economic activities are concentrated geographically. Most people in advanced countries, and a growing number in developing countries, live in large, densely populated metropolitan areas. Many industries—including service industries such as banking—also are concentrated geographically, and such clusters are an important source of international specialization and trade. Yet we do not all live in one big city, nor does the world economy concentrate production of each good in a single location. Why?

CENTRIPETAL AND CENTRIFUGAL FORCES

Obviously there is a tug of war between forces that promote geographic concentration and those that oppose it—between centripetal and centrifugal forces. These forces can be represented by the items shown in Table 1. This list is not comprehensive; it is merely a selection of some forces that may be important in practice.

The centripetal forces listed in the first column of Table 1 are the three classic Marshallian sources of external economies. A large local market creates both backward linkages (i.e., sites with good access to large markets are preferred locations for the production of goods subject to economies of scale) and forward linkages (i.e., a large local market supports the local production of intermediate goods, lowering costs for downstream producers). An industrial concentration supports a thick local labor market, especially for specialized skills, so it is easier for workers to find employers and for employers to find workers. And a local concentration of economic activity may create more or less pure external economies through information spillovers.

The centrifugal forces in the second column of Table 1 are less standard but offer a useful breakdown. Immobile factors, certainly land and natural resources and, in an international context, people militate against concentration of production, both from the supply side (some production must go to where workers are) and from the demand side (dispersed factors create a dispersed market, and some production will have an incentive to locate close to consumers). Concentrations of economic activity increase the demand for local land, driving up land rents and so discouraging further concentration. And concentrations of activity can generate more or less pure external diseconomies such as congestion.

In the real world, agglomeration in general, as well as any example of it, typically reflects all of the items in Table 1. Why is the financial services industry concentrated in New York City? Partly because the city's size makes it an attractive place to do business and because the concentration of the financial industry means that many clients and ancillary services are located there. Also important are the city's thick market for those with special skills, such as securities lawyers, and the general importance of being in the midst of the buzz. But why isn't all financial business concentrated in New York City? Partly because many clients are not there, partly because office space is expensive, and partly because it is a nuisance to deal with the city's traffic, crime, and other urban realities.

To conduct analytical work on economic geography, however, it is necessary to cut through the complexities of the real world and focus on a more limited set of forces. In fact, the natural thing is to pick one force from the first column of Table 1 and one from the second: to focus on the tension between just one centripetal and one centrifugal force. In the line of work on economic geography started by my 1991 article and book (Krugman 1991a, 1991b), most models have chosen the first item in each column, analyzing linkages as the force for concentration and immobile factors as the force opposing concentration.

These choices are dictated less by empirical judgment than by two strategic modeling considerations. First, it is desirable to put some distance between assumptions and conclusions—to avoid an approach that appears to assert that agglomeration takes place because of agglomeration economies. Much of the analysis we want to undertake involves asking how a changing economic environment alters economic geography. This will be an ill-defined task if the forces producing that geography are inside a black box labeled external effects. So the pure external economies and diseconomies are put to one side, in favor of forces that are more amenable to analysis.

Second, if location is the issue, it is helpful to be able to deal with models in which distance enters in a natural way. Linkage effects, which are mediated by transportation costs, are naturally tied to distance—so is access to immobile factors. By contrast, the thickness of the labor market must have something to do with distance, but it does not lend itself quite so easily to being placed in a spatial setting. And land rents as a centrifugal force pose conceptual challenges—the "infinite Los Angeles problem"—that are discussed briefly in the section on chance and necessity below.

MODELING TRICKS

The idea is hardly new that there may be a circular process in which the decisions of individual producers to choose a location with good access to markets and suppliers improve the market or supply access of other producers in that location. Indeed, that was the central theme of studies by Harris (1954) and Pred (1966), both well-known among geographers. Why, then, did this idea not become widely known in economics until the 1990s?

The most likely answer is that underlying the work of Harris and Pred is the implicit assumption that there are substantial economies of scale at the level of the plant. In the absence of such scale economies, producers would have no incentive to concentrate their activity: they would simply supply consumers from many local plants. An expansion of a regional market would not predictably lead to an increase in the range of goods produced in that region. Increasing returns, in other words, are central to the story.

The same may be said of spatial economics in general. Almost all of the interesting ideas in location theory rely implicitly or explicitly on the assumption that important economies of scale enforce the geographic concentration of some activities. Thus, Weber's (1909) analysis of the location decisions of an individual producer trying to minimize the combined costs of production and delivery assumes that there can be only one production site; Christaller's (1933) suggestion that cities form a hierarchy of central places depends on the assumption that larger cities can support a wider range of activities, and Lösch's (1940) famous demonstration that an efficient pattern of central places would imply hexagonal market areas assumes that some economic activities can be undertaken only at a limited number of sites. (The main example of a location model that does not rely on some form of scale economies, the land-rent analysis of von Thünen [1826], in effect hides the role of increasing returns by simply assuming the existence of a central city.) But unexhausted economies of scale at the level of the firm necessarily undermine perfect competition.

The reason geography has finally made it into the economic mainstream is therefore obvious: imperfect competition is no longer regarded as impossible to model, so stories that crucially involve unexhausted scale economies are no longer out of bounds. Indeed, the new interest in geography can be viewed as the fourth (and final?) wave of the increasing returns-imperfect competition revolution that has swept through economics over the past twenty years. First came the new industrial organization, which created a toolbox of tractable if not entirely convincing models of imperfect competition. Then the new trade theory, which used that toolbox to build models of international trade in the presence of increasing returns. Then the new growth theory, which did much the same for economic growth. What happened after 1990 was the emergence of the new economic geography, which might best be described as a genre of economic analysis that tries to explain the spatial structure of the economy using technical tricks to produce models in which there are increasing returns and markets characterized by imperfect competition. Fujita, Krugman, and Venables (forthcoming) summarize these tricks as results of "Dixit-Stiglitz, icebergs, evolution, and the computer." Why, and how?

DIXIT-STIGLITZ

The remarkable model of monopolistic competition developed by Dixit and Stiglitz (1977) has become a workhorse in many areas of economics. In the new economic geography it has one especially appealing feature: because it assumes a continuum of goods, it lets modelers respect the integer nature of many location decisions—no fractional plants allowed—yet analyze their models in terms of the behavior of continuous variables such as the share of manufacturing in a particular region. In effect, Dixit-Stiglitz lets us have our cake and cut it into arbitrarily small pieces too.

ICEBERGS

Icebergs are a less familiar technical trick. Transportation costs are of the essence in the new economic geography. Yet any attempt to develop a general equilibrium model of economic geography would be substantially complicated by the need to model transportation as well as goods-producing sectors. Worse yet, transportation costs can undermine the constant demand elasticity that is one of the crucial simplifying assumptions of the Dixit-Stiglitz model. Both problems can be sidestepped with an assumption first introduced by Samuelson (1954) in international trade theory: a fraction of any shipped good simply "melts away" in transit so that transport costs are in effect incurred in the good shipped. (In new economic geography models, melting is usually assumed to take place at a constant rate per distance covered, for example, 1 percent of the cargo melts away per mile.) In terms of modeling convenience, there turns out to be a spectacular synergy between the Dixit-Stiglitz market structure and iceberg transport costs: not only can one avoid the need to model an additional industry but because the transport cost between any two locations is always a constant fraction of the free on board (f.o.b.) price, the constant elasticity of demand is preserved.

Evolution

Interesting stories about economic geography often seem to imply multiple equilibria. Suppose, for example, that producers want to locate where other producers choose to locate; this immediately suggests some arbitrariness about where they actually end up. But which equilibrium does the economy select? New economic geography models typically assume an ad hoc process of adjustment in which factors of production gradually move toward locations that offer higher current real returns. This sort of dynamic process was initially proposed apologetically because it neglects the role of expectations. But it is possible to regard models of geography as games in which actors choose locations rather than strategies—or rather, in which locations are strategies—in which case one is engaged not in old-fashioned static expectations analysis but rather in state-of-the-art evolutionary game theory! (To middle-brow modelers such as myself it sometimes seems that the main contribution of evolutionary game theory has been to relegitimize those little arrows that we always want to draw on our diagrams.)

The Computer

Finally, despite the best efforts of theorists, all but the simplest models of economic geography usually turn out to be beyond the reach of paper-and-pencil analysis. As a result, the genre relies to an unusual extent on numerical examples—on the exploration of models using both static calculations and dynamic simulations.

Dynamics of Geographic Change

Suppose that an economic activity has a slightly larger initial concentration in one location than in another. Will that concentration be self-reinforcing, with a growing disparity between the locations, or will there be a tendency back toward a symmetric state? The answer presumably depends on the relative strength of centripetal and centrifugal forces.

Suppose, on the other hand, that a concentration of economic activity already exists but that some of that activity moves elsewhere. Will the activity move back, or will the concentration unravel? The answer to this question similarly depends on the relative strength of centripetal and centrifugal forces.

As these generic questions suggest, models of economic geography typically exhibit a pattern in which the qualitative behavior of the model changes abruptly when the quantitative balance of forces passes some critical level. That is, the

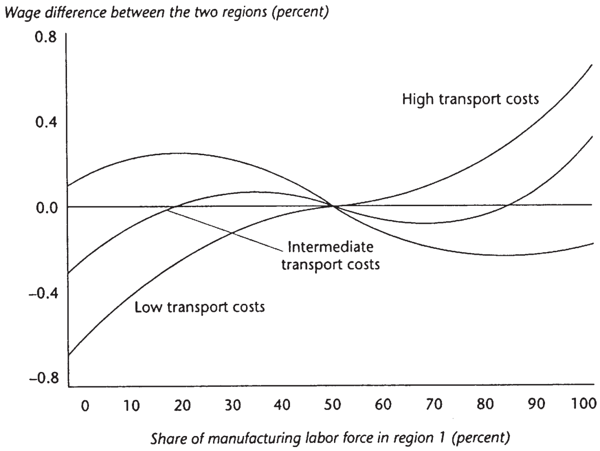

FIGURE 1. Relationship between Regional Manufacturing Populations and Real Wages with Varying Transport Costs

models are characterized by bifurcations. And bifurcation diagrams are therefore a central analytical tool in this literature.

The typical forms of these bifurcations are illustrated in Figures 1 and 2, which show results from a simulation of the model introduced in Krugman (1991b). That article was, in effect, an attempt to formalize the story suggested by Harris (1954) and Pred (1966). The model envisaged an economy consisting of two regions, each with two industries: immobile, perfectly competitive agriculture and mobile, imperfectly competitive (Dixit-Stiglitz) manufacturing. The backward and forward linkages in manufacturing generated centripetal forces; the pull of the immobile farmers generated the centrifugal force.

Figure 1 shows how the difference in real wages between the two regions depends on the allocation of manufacturing between them (a calculation that involves repeatedly solving a small computable general equilibrium model). The horizontal axis shows the share of manufacturing workers living in region 1; the vertical axis shows the difference between real wages in region 1 and region 2. Each curve is calculated for a different level of transport costs.

The rough intuition behind these curves runs as follows. If transport costs are high, there is relatively little interregional trade. So the wages that workers can earn depend mainly on the amount of local competition and thus decrease as the number of other workers in the same region increases. When transport costs are low, a typical firm sells extensively in both regions. But because it has better access to markets if it is located in the region with the larger population of workers, it can afford to pay higher wages, and the purchasing power of those wages also is higher because

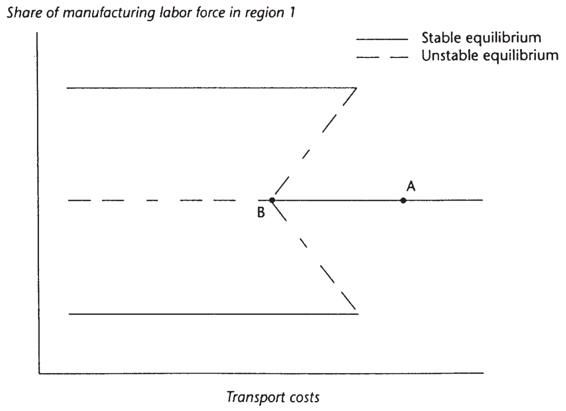

FIGURE 2. Relationship between Regional Manufacturing Populations and Transport Costs

workers have better access to consumer goods. So in that case, real wages increase with a region's population of workers. At intermediate transport costs, these two forces are nearly balanced. The particular curve shown, in which centripetal forces are stronger when regions are very unequal, whereas centrifugal forces are stronger when they are nearly symmetric, is an artifact of the particular functional forms used in this exercise.

Because workers are assumed to move to whichever region offers the higher real wage, in the case of high transport costs there is a unique equilibrium with workers evenly divided between the regions. In the case of low transport costs there are three equilibria—one with workers evenly divided and two with workers concentrated in either region. And in the intermediate case, there are five equilibria.

Figure 2 shows the bifurcation diagram that results from the assumption that workers gradually move toward the region offering the higher real wage. It shows how the set of equilibria—as measured by the share of the manufacturing labor force in region 1—depend on transport costs, with solid lines indicating stable and broken lines indicating unstable equilibria. The figure illustrates nicely one of the appealing features of the new economic geography: it easily allows one to work through interesting "imaginary histories." Suppose, for example, that we imagine an economy that starts with high transport costs and therefore with an even division of manufacturing between regions, a situation illustrated by point A in Figure 2. Then suppose that transport costs were to fall. When the economy reached point B, it would begin a cumulative process in which a growing concentration of manufacturing in one region would lead to an ever-larger concentration of manufacturing in that region. That is, the economy would spontaneously organize itself into a core-periphery geography.

Geographic Theories of the World Economy

A generation ago it was common for critics of the economic system to argue that developing countries were not simply economies on the same road as industrial economies, although less advanced. Rather, they argued, the emergence of rich and poor countries was part of a common process of uneven development in which initial advantages in certain regions had accumulated over time, giving them a privileged economic position while relegating the rest of the world to a subordinate role as hewers of wood and drawers of water. In the past ten years, worries have largely reversed: advanced countries seem to fear that newly industrializing economies will undermine the North’s prosperity.

The Theoretical World, with Two Economies

New economic geography models can shed light on both concerns. The models suggest that both the differentiation of the world into high-wage industrial core and low-wage nonindustrial periphery, and a subsequent period of dispersal of industry and convergence of wages, can be explained by an ongoing process of declining trade costs.

The basic concepts were introduced by Venables (1995). He assumed, in contrast to the regional model described in the previous section, that factors were completely immobile between countries. However, a possibility for cumulative processes was introduced by making a distinction between a constant-returns agricultural sector and an increasing-returns manufacturing sector that both uses and produces intermediate inputs. The basic idea is that intermediate goods producers in a region with a large manufacturing sector will have superior access to the large markets afforded by downstream producers (backward linkage), whereas these producers in turn will have the advantage of better access to the intermediate goods produced in their own region (forward linkage). In the original formulation, the upstream and downstream components of manufacturing were treated as separate sectors; in subsequent work, including Krugman and Venables (1995) and Puga and Venables (1997), the same differentiated goods were assumed to enter into consumption and production, allowing a consolidation of the sector into a common manufacturing aggregate.

Suppose now that we imagine a world consisting of two initially identical regions, with varying costs of transporting manufactured goods between them. If transport costs are high, each region will essentially be self-sufficient and the regions will therefore be symmetric in outcomes as well as initial conditions. But now imagine gradually falling transport costs. It now becomes increasingly possible for firms to export their manufactured goods to the other region. Yet because of transport costs, production in whichever region has the larger manufacturing sector (because of any small difference or simple historical accident) will benefit from better access to both markets and suppliers. Thus, when transport costs drop below some critical level, a process of differentiation between regions will take place, with manufacturing concentrating in a core while the periphery is relegated to primary production.

The impact of this process depends on the size of the manufacturing sector, more specifically, on the share of manufactured goods in spending. If this share is low, the region that becomes the core does not get a significantly higher wage rate from that role. But if the share is sufficiently large (in a two-region model, if it exceeds half of total spending on traded goods), the core ends up with higher wages than the periphery, and the process of differentiation can be immiserizing for the peripheral region. This simple approach, then, offers a possible justification for claims that the backwardness of the South is not something that developed in isolation: it is a necessary consequence of the process that also produced the industrialization of the North.

Perhaps more surprisingly, the same model predicts that a continuing decline in transport costs—loosely speaking, the continuing process of globalization—eventually produces a reversal of fortune. The reason is that the peripheral region has a competitive advantage in the form of lower wages. At first this advantage is more than offset by the North's superior access to markets (backward linkage) and inputs (forward linkage). But as transport costs fall, the importance of these linkages also declines. So there is a second critical point at which industry finds it profitable to move to lower-wage locations.

This is a surprisingly satisfying result: by imagining a hypothetical history in which a single driving variable—transport costs—follows a monotonic path through time, we are able to derive an evolutionary path for the world economy in which the inequality of nations and the division of the world into primary and industrial producers first spontaneously emerges, then dissolves. Understandably, then, Venables and I referred to the original article as the "history of the world, part I."

THE REAL WORLD, WITH MANY ECONOMIES

I will return shortly to the question of how much of the history of the real world such an analysis actually captures. First, however, it is useful to use the geographic theories of the world economy as an occasion to discuss the spatial aspects of modeling.

The analysis in Krugman and Venables (1995), like much international trade theory, imagines a world with just two discrete locations, themselves modeled as points. It involves space only to the extent that there are assumed to be transport costs between these points. To a serious geographer, of course, this is grossly inadequate: the spatial relationships both between and within countries should be taken into account. Indeed, as a first approximation, a geographer might even want to ignore national boundaries, asking how an undifferentiated, seamless world economy might develop a spatial structure.

To do this in general is probably impossible. Indeed, as soon as one goes even a bit beyond a two- or three-location world, the whole exercise tends to bog down in uninformative taxonomy. But it is possible to gain considerable insight by focusing on particular, unrealistic, but convenient geometries for the world.

One particular geometry that is useful despite its artificiality is what we might call the "racetrack" economy: a large number of regions located symmetrically around a circle, with transportation possible only around the circumference of that circle. This setup has two useful properties. First, the economy is one-dimensional, which greatly simplifies both algebra and calculations. Second, because there are no edges and hence no center, it is a convenient way to retain the feature that all sites are identical, which means that any spatial structure that emerges represents pure self-organization.

If one takes a racetrack version of the Krugman and Venables (1997) model and starts it with an almost but not quite uniform distribution of manufacturing across space, what happens is a spontaneous differentiation into manufacturing and agricultural regions. The size and spacing of these regions are predictable, even if the initial deviation from uniformity is random. The reason for this predictability was, it turns out, explained in a seemingly different context—morphogenesis in theoretical biology—by, of all people, Alan Turing (1952). But the question of which parts of the world take on which role remains arbitrary, a function of small initial advantages that determine the phase of the regional development pattern (i.e., how the alternating bands of industry and agriculture are rotated around the circle).

Extending this sort of analysis to more realistic geometries turns out to be startlingly difficult. Still, the racetrack analysis is at least suggestive of the reasons that patterns of development and underdevelopment are regional—why, for example, all of northwestern Europe shared in the industrial revolution—rather than confined within national boundaries.

What about the larger story of the rise and fall of international inequality? Surely the forces covered in this approach do not tell the full story, or perhaps even more than a small part of the real story. In particular, if one tries to put realistic shares of North-South trade in gross world product into the model, it is difficult to make either the initial divergence of incomes as the world divides itself into industrial and primary-producing regions, or the later spread of industry, have impacts on real income in either region of more than a few percent. There may be ways to make the story take on greater significance—say, by introducing some interaction between patterns of trade specialization and external economies in domestic production. But at this point it would be premature to take the interesting and suggestive "history of the world" as more than a possible story about part of what actually happened.

REGIONAL INEQUALITY IN DEVELOPING COUNTRIES

It is often observed that many developing countries suffer from significant economic dualism, in which a relatively high-wage, high-income economy appears to exist within a much less developed economy, and that this dualism has a strong geographic dimension. Although a lot of development economics continues to treat countries as dimensionless points, in other contexts the contrast between Mexico City and Chiapas, or between São Paulo and Brazil's northeast, looms large.

It is not difficult to convert the core-periphery analysis discussed above into a story of regional divergence. One need only relabel the workers of that model with mobile factors, such as capital and skilled labor, and presume that unskilled labor is a (relatively) immobile factor, so that it takes on the role of the farmers. The story can be made more realistic, adding complications but no essential differences, by allowing the mobile and immobile factors to be substitutes in production. With sufficiently strong scale economies and transport costs, the resulting core-periphery equilibrium can have large wage differentials for the immobile factor.

In words, this story says that Brazil's south is a more attractive place to produce than its north because of the concentration of purchasing power and availability of intermediate inputs in the south and that because of this attraction those factors of production that can move have concentrated in the south, sustaining the concentration of markets and suppliers that creates the south's advantage. As in all the models discussed in this article, the original source of the south's advantage need not lie in any inherent superiority of its resources or location: it could simply be the result of historical accident.

Although this is a coherent story, some modeling of regional inequalities has suggested an additional source of those inequalities: self-reinforcing advantages of market access through transportation networks. One simple version of this story was laid out in Krugman (1993) and is illustrated in the left side of Figure 3. The figure shows three locations; the width of the lines between the locations is an inverse indicator of transport costs (i.e., thicker lines mean lower costs, just as thicker means better on a road map). As drawn, location 1 is obviously a transport hub in the sense that it is cheaper to get from location 1 to either of the other locations than it is to go between those locations. It is easy to show that other things being equal (i.e., given the same market sizes and availability of locally produced inputs) this will make location 1 more attractive for producers subject to increasing returns. So a transport hub will be a favored location for industry. (Like many observations in the new economic geography, this is a painfully obvious point that somehow just was not in the literature before.)

But why should transport costs be lower between location 1 and other locations than between those other locations? One obvious answer is that if industry is

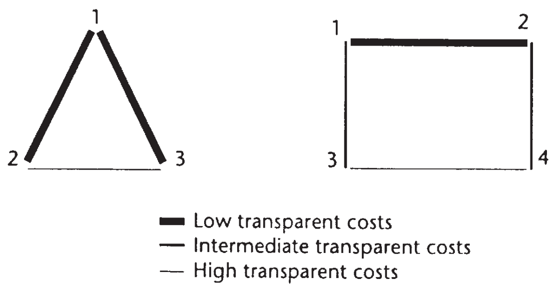

FIGURE 3. Transparent Costs, Transparent Hubs, and Industrial Concentration

concentrated in location 1, there will be more trade between location 2 and location 1 than between location 2 and location 3, and so on. And if there are increasing returns in transportation—as there surely are—this will mean lower per unit transport costs along the more heavily used routes.

Clearly, we have another example here of a self-reinforcing process: a location that for whatever reason has a concentration of production will tend to become central in terms of the transport network, which will reinforce its advantage as a production location, and so on. Krugman (1993) shows that this process can produce a core-periphery pattern of industrialization even if we suppress the factor mobility that drives the standard models of such patterns.

A slightly different role for favored transport access is illustrated on the right side of Figure 3. Here we see four locations, with transport costs lower between location 1 and location 2 than between either of those locations and the rest of the economy and with transport costs between locations 3 and 4 particularly high. This pattern might emerge, again in the presence of increasing returns in transportation, if locations 1 and 2 both had large concentrations of industry. The effect, of course, would be to make locations 1 and 2 more attractive places to do business, reinforcing their advantage. A concrete example: part of São Paulo's advantage is its good access to Rio de Janeiro, including frequent plane flights, and vice versa. This is natural between Brazil's two largest cities but further reinforces the tendency of activity to concentrate in those two cities.

Just as in the global economy models discussed earlier, models of regional inequality can easily show a nonmonotonic response to declining transport costs. Initially, such declines can promote the formation of core-periphery patterns. To take a classic example, the stark division of Italy into affluent north and less affluent south took shape when railroads were introduced. Railroads made it possible for factories in the north to supply the needs of agricultural markets in the south, causing deindustrialization in the south. Moreover, the railroad network did more to connect the already industrialized regions of the north than those of the south, reinforcing the advantage of those northern locations in terms of access to markets and inputs.

Eventually, however, sufficiently low transport costs (even on a small scale of transportation) can lead to a spread of industry: once it is inexpensive to transport inputs wherever they are needed and export products from any location, the lower factor costs of the periphery become increasingly significant. (In Brazil, there is currently some relocation of industry to the northeast, where wages are about one-third the levels in São Paulo. This is one of the factors often blamed for rising unemployment in traditional industrial areas.) Of course, regional inequality may also be strongly influenced by government policy-including trade policy, as described below.

Policy and Primacy

A striking feature of many developing countries is the existence of one huge urban concentration, normally the capital city. Why are urban giants in developing countries so large?

Empirical studies of primacy identify two strong factors determining the size of the largest city: urban population as a whole and, more interestingly, political structure—primary cities are smaller in federal and decentralized systems than in highly centralized systems. Thus, Mexico City is still larger than Shanghai because of China's decentralization.

The role of political centralization in primacy is fairly obvious at one level: it results from the direct demand and employment created by the government apparatus and from the more subtle advantages of access to government officials. (When one asks Japanese executives why they are willing to pay the high cost of keeping their headquarters in central Tokyo, access to officials is usually the first thing they mention.)

The type of analysis described in this survey suggests, however, that beyond these direct effects one might well expect a multiplier effect, perhaps even a catalytic effect, of political centralization (see the section on geography and policy, below). That is, whatever initial concentration of demand and advantages of access are conveyed to businesses in the capital will be magnified through the usual circular processes involving market size, access to suppliers, transportation advantages, and so on. Such magnification effects may explain the extraordinary strength of the relationship between political centrality and primacy (e.g., the fact that Tokyo is substantially larger than New York even though Japan has only half as many people as the United States).

There also may be other important policy linkages. Hanson (1992) notes that Mexico's trade liberalization in the late 1980s seemed to be associated with a dramatic decentralization of manufacturing away from Mexico City—not only with the growth of new export centers near the U.S. border but also a spinning out of industries producing for the domestic market. In Krugman and Livas Elizondo (1996), an effort was made to justify this observation in terms of a formal model. The article envisaged a domestic economy with two locations and mobile manufacturing; the necessary centripetal force was supplied by backward and forward linkages, the centrifugal force by land rents. However, these two locations were assumed to trade (but not have factor mobility) with a large third region, the rest of the world.

The point we then made was that the importance of the linkages supporting population concentration within this country would depend on its trade policy. Suppose that the country was strongly protectionist and hence did little external trade. Then domestic producers would mainly sell to domestic consumers and buy inputs from other domestic producers. The result would be strong linkage effects that would tend to promote and sustain a concentration of manufacturing in only one location. But if trade were liberalized, domestic producers would sell much of their output abroad—and hence have less incentive to locate near the large domestic market—and would also buy many of their inputs from abroad—and hence have less incentive to locate near domestic suppliers. Meanwhile, high land rents would still create an incentive to locate away from other producers. Numerical examples confirm that high trade barriers would tend to foster concentration of manufacturing in a single Mexico City–type location, whereas reduced trade barriers would tend to cause such concentrations to unravel. (An interesting question would be whether Brazil's trade liberalization has similarly contributed to the apparent shift of manufacturing away from its traditional centers in the south. If so, it would be a cleaner example of our story than the case of Mexico because proximity to the border is not an issue—indeed, given the Mercosur trade union, the border issue actually cuts the other way.)

For what it is worth, cross-sectional regressions by Ades and Glaser (1995) find evidence that inward-looking trade policies foster the creation of urban giants, although other factors appear to be more important. However, one may question whether the highly nonlinear stories told by the models can be tested very well by such regressions. (Empirical work in this area is generally difficult for that reason.)

Chance and Necessity

At the beginning of this article I described two approaches that both go under the rubric of geography but seem to take diametrically opposed positions: the type of model described above, in which there are multiple equilibria and the geographical pattern of production depends on historical accidents, and the approach recently promoted by John Luke Gallup, Jeffrey Sachs, and Andrew Mellinger (see elsewhere in this issue), in which differences in natural geography exert powerful influences on economic development. But I also suggested that this may be a false dichotomy.

To illustrate this point, consider Mexico City. The concentration of population and production in the Valley of Mexico has deep historical roots, essentially environmental in nature: before the Spanish conquest, the Aztecs practiced a highly productive form of agriculture made possible by the existence of a large lake, which supported a dense local population (by preindustrial standards). It was natural that this location should become the site for Mexico's main urban center. But the valley no longer contains a lake, or for that matter any agriculture to speak of. Today, Mexico City is there because it is there, its existence sustained by the kinds of circular processes discussed in earlier sections. So in one sense, the location of Mexico's primary city was dictated by natural geography. Yet those geographic advantages are no longer relevant in any direct sense, and they have been able to cast such a long shadow over the future only because the geography of the economy has such strong self-reinforcing features that a concentration of population, once established, tends to persist and even grow. (The role of the Erie Canal in giving New York City its dominant position is a classic first-world example of the same proposition.)

Put another way, in many cases, aspects of natural geography matter a lot not because natural features of the landscape are crucial but because they inspire self-reinforcing agglomerations. So it is precisely the aspects of the economy that in principle allow history-dependent, multiple-equilibria stories to be told that in practice give exogenous geography such a strong role.

In formal models of economic geography, especially when one allows the geography of the economy to evolve over time, it often turns out that small nonhomogeneities in the landscape have dramatic effects on the outcome. Thus, in the core-periphery models of the first two sections, giving one of the regions a small advantage in the size of its agricultural base removes the arbitrariness of which region will become the core and which the periphery as transport costs fall below the critical level. This means that a small difference in inherent advantage can produce a large difference in outcomes. (It also turns out that small inherent differences strongly bias the outcome when one starts with some random allocation of mobile factors.)

Most recent work making this point has concentrated on the effect of natural differences in transport costs on urban location, explaining why, for example, most great cities are ports, even though in the modern world few large cities derive much of their income or employment from that role (Fujita and Mori 1996a).

It is possible to imagine a variant on the models developed earlier in which all factors except land are mobile. In such a model it is possible, provided the economy is not too large, to have a self-sustaining "von Thünen" spatial pattern in which manufacturing is concentrated at a single location surrounded by an agricultural hinterland. However, if one imagines a gradually increasing population, eventually it becomes profitable for some manufacturing to locate away from the original center and new cities emerge.

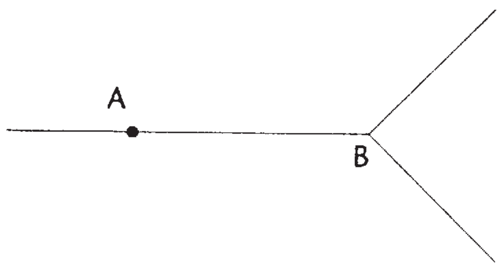

FIGURE 4. Formation and Location of New Cities

But where do they emerge? Figure 4 represents a version of Fujita and Mori's analysis. Suppose the economy is in a long, narrow valley (making it effectively one-dimensional), with the original city at location A. If a fork is put in the valley at location B, it is effectively a point with superior access to the rest of the world than other locations, which makes it a stylized representation not only of the role of a river or road junction but of a port as well.

As the population expands, location A's agricultural hinterland will expand as well, eventually pushing up both forks of the valley beyond location B. And eventually a new city will emerge. Where? Location B is a very likely location in the following sense: imagine choosing alternative initial positions for location A (or varying any other parameter of the model) and asking where the next city will emerge. In general, any possible location will be chosen for at most one location of the original city. But because of the special advantages of location B (which turn out to generate a cusp in the market-potential function that determines location choice), there is a nonzero-length range of initial city locations that will lead the second city to emerge there. So natural geography will often (although not always) dictate the city site. Yet once the city is established, those natural advantages will be much less important a reason for the "lock-in" of its location than the self-sustaining advantages of an established concentration of activity.

The paradox that natural geography may matter so much precisely because of strong circular causation has important implications for the interpretation of correlations between natural advantages and actual economic geography. These correlations may say more about the processes that have produced the geography we see than about what might be possible in the future. To take the Fujita-Mori analysis as an example: the historical role of ports as sites around which cities crystallize explains why most of today's large cities are ports. But because the importance of the port was only that of serving as a springboard, and is not a major current source of advantage, it need not be the case that future cities also be ports. If, say, an inland city were constructed as a deliberate national policy and supported effectively, it might become self-sustaining even though its location does not fit any of the criteria that characterize today's major cities.

To put a sharper point on it: the current pattern of world economic geography shows a strong association between per capita income and essentially Western European conditions—temperate climate, absence of malaria, much of the population close to the coast or navigable rivers, and so forth. But this pattern may mainly reflect the catalytic role of these factors in the past and need not imply that an inland country (which now has access to good roads and cheap air transport) with a hot climate (but now has access to modern cooling technology) and environmental conditions that once made it malarial (but not now thanks to mosquito eradication programs) cannot break free of its low-level trap and move to a better equilibrium. All of this brings us to policy.

GEOGRAPHY AND POLICY

This will, necessarily, be a short conclusion. At this point, little effort has been made to draw policy conclusions from the new economic geography literature. The main goal for the moment is to explain why.

In principle, the sort of economy envisaged by the models sketched out in this article should be a prime target for government intervention. There is no presumption here that the market will get it right. Moreover, the models suggest that under some circumstances, small policy interventions can have large and perhaps lasting effects. Finally, because cumulative processes of concentration tend to produce winners and losers, perhaps at the level of nations, there is an obvious incentive for policy makers to try to make sure that their nation emerges as one of the winners.

Nonetheless, those of us working on these models have been extremely cautious about drawing policy implications. Mainly this reflects a strong sense of how difficult it is to go from suggestive small models to empirically based models that can be used to evaluate specific policies. The long debate over the applicability of the theory of strategic trade policy, which eventually led mainly to an appreciation of just how hard it is to map reality into even sophisticated models of imperfect markets, is fresh in the minds of many of the relevant theorists. And new geography models, in which the crucial effects are general equilibrium rather than merely partial equilibrium, are likely to be even harder to make operational.

There also is, to be honest, concern (at least on my part) that some of the less pleasant aspects of the history of strategic trade policy will be repeated: the frantic efforts of interested parties to recruit reputable economists to endorse questionable interventionist policies. Admittedly, that temptation was admirably resisted by all the major players in the new trade theory, but it was not an experience one wants to encourage.

But there also is a special consideration that makes policy conclusions difficult in the geographic literature. Consider Table 1 again, bearing in mind that in most cases all the entries will be relevant. What is immediately striking is that there are external effects on both sides. So there is a market failure case to be made both that any given agglomeration is too big (look at the congestion and pollution) and too small (think of the linkages and spillovers that would come with more activity). One may have opinions—I am quite sure in my gut, and even more so in my lungs, that Mexico City is too big—but gut feelings are not a sound basis for policy.

One recommendation is safe, however. Because geography is such a crucial factor in development, and there are undoubtedly strong policy implications of some sort, it is an important subject for further research.

REFERENCES

Ades, A., and E. Glaser. 1995. Trade and circuses: Explaining urban giants. *Quarterly Journal of Economics* 110: 195-227.

Christaller, W. 1933. *Central places in southern Germany.* Jena, Germany: Fischer Verlag.

Davis, D., and R. Weinstein. 1997. *Empirical testing of economic geography: Evidence from regional data*. Cambridge, MA: Harvard University Press.

Dicken, P., and P. Lloyd. 1990. *Location in space: Theoretical perspectives in economic geography.* New York: HarperCollins.

Dixit, Auinash, and Joseph E. Stiglitz. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67: 297-308.

Fujita, Masahisa, and Paul Krugman. 1995. When is the economy monocentric: von Thünen and Christaller unified. *Regional Studies and Urban Economics* 25 (August): 505-28.

Fujita, Masahisa, Paul Krugman, and Anthony Venables. Forthcoming. *The spatial economy.*

Fujita, Masahisa, and Tomoya Mori. 1996a. The role of ports in the making of major cities: Self-agglomeration and hub-effect. *Journal of Development Economics* 49 (April): 93-120.

———. 1996b. Structural stability and evolution of urban systems. Regional Science and Urban *Economics.*

Fujita, Masahisa, Tomoya Mori, and Paul Krugman. 1999. On the evolution of hierarchical urban systems. *European Economic Review* 43 (February): 209-51.

Harris, C. D. 1954. The market as a factor in the localization of production. *Annals of the Association of American Geographers* 44: 315-48.

Hoover, Edgar M., and Raymond Vernon. 1959. *Anatomy of a metropolis*. Cambridge, MA: Harvard University Press.

Karaska, G., and D. Bramhall, eds. 1969. *Locational analysis for manufacturing*. Cambridge, MA: MIT Press.

Krugman, Paul. 1991a. *Geography and trade*. Cambridge, MA: MIT Press.

———. 1991b. Increasing returns and economic geography. *Journal of Political Economy.*

———. 1993. On the number and location of cities. *European Economic Review* 37 (April): 293-98.

Krugman, Paul, and R. Livas Elizondo. 1996. Trade policy and the third world metropolis. *Journal of Development Economics* 49: 137-50.

Krugman, Paul, and Anthony Venables. 1995. Globalization and the inequality of nations. *Quarterly Journal of Economics* 110: 857-80.

———. 1997. *The seamless world: A spatial model of international specialization and trade*. Cambridge, MA: MIT Press.

Lösch, August. 1940. *The economics of location*. Jena, Germany: Fischer Verlag.

Pred, A. R. 1966. *The spatial dynamics of U.S. urban-industrial growth, 1800-1914*. Cambridge, MA: MIT Press.

Puga, D., and Anthony Venables. 1997. The spread of industry: Spatial agglomeration in economic development. Centre for Economic Policy Research working paper 1354, London.

Samuelson, Paul. 1954. The transfer problem and transport costs. *Economic Journal* 64 (254): 264-89.

Thünen, J. von. 1826. *The isolated state*. London: Pergamon.

Turing, Alan. 1952. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London* 237: 37.

Venables, Anthony. 1995. Equilibrium locations of vertically linked industries. *International Economic Review.*

Weber, A. 1909. *Theory of the location of industries*. Chicago: University of Chicago Press.

Weibull, M. 1995. *Evolutionary game theory.* Cambridge, MA: MIT Press.