The Role of Geography in Development

Paul Krugman

April 1998

Paper prepared for the Annual World Bank Conference on Development Economics, Washington, D.C., April 20–21, 1998.

The Role of Geography in Development

Paul Krugman

Abstract

The recent surge of interest in the role of economic geography in economic development has divided into two seemingly contradictory approaches. One approach emphasizes the role of inherent features of the landscape in shaping development patterns. This paper, however, mainly surveys the alternative approach, which stresses 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 more or less symmetric locations can end up playing very different economic roles.

Such processes can occur at several different levels. The paper discusses "geographical" models of the division of the world into industrial and nonindustrial countries, of the emergence of regional inequality within developing countries, and of the emergence of giant urban centers. The paper also argues that the conflict between "self-organizing" and "predestination" approaches to economic geography may be more apparent than real: natural features matter so much largely because they act as seeds around which cumulative processes crystallize, so that while geography may have been destiny in the past, it need not be in the future.

Finally, the paper discusses policy briefly, mainly in terms of why it is so hard to draw policy conclusions from these models.

The Role of Geography in Development

Paul Krugman

In recent years there has been a surge of interest in the geographical aspects of development—that is, in the question of *where* economic activities take place. There is nothing surprising about this interest—or to put it differently, perhaps the surprise is that it took so long for this interest to become a main stream concern within economics. After all, even a casual look at the map 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 within Europe seems to follow a downward gradient from the northwest corner of the continent. It is also apparent that there are both large regional inequalities in development within countries and, often, a powerful tendency for populations to concentrate in a few densely populated regions (and as any traveler can tell you, in a few huge cities). But only recently have the attempts to explain such locational patterns become a subject for research by large numbers of economists.

Within the new interest in economic geography, there is what may seem to be a paradoxical difference between two general approaches. One approach—the approach exemplified by Jeffrey Sachs's paper for this conference—attempts to explain the differences in economic development in varying 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. 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 10-million-person metropolitan nightmare. These two approaches may well seem to be contradictory: one seems to be a story of predetermination, the other a story of the capricious role of chance. As I will argue later in this paper, however, the contradiction is more apparent than real: in fact, understanding why small random events can produce large consequences for economic geography is also 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, the bulk of this paper is devoted to understanding the ways in which 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 nonetheless end up playing very different economic roles.

The New Economic Geography: Theoretical Principles

Petals versus Fugals

Many economic activities are markedly 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) are also

geographically concentrated, and such clusters are clearly 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. Obviously there is a tug of war between forces that tend to promote geographical concentration and those that tend to oppose it—between "centripetal" and "centrifugal" forces. (Since these terms, while natural and useful, are very difficult to distinguish when pronounced, I find it useful to refer to them in short as "petals" versus "fugals.")

What are these forces? We may represent them in terms of a menu of the type shown in table 1. The menu should not be viewed as comprehensive; it is a selection of *some* forces that may be important in practice. It shows two columns: one of centripetal forces, one of centrifugal forces.

Table 1. Forces Affecting Geographical Concentration

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

The petals listed on the left side of table 1 are the three classic Marshallian sources of external economies. A large local market creates both "backward linkages"—that is, sites with good access to large markets are preferred locations for the production of goods subject to economies of scale—and "forward linkages"—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 that employees find it easier to find employers and vice versa. And a local concentration of economic activity may

create more or less pure external economies through information spillovers. ("The mysteries of the trade become no mystery, but are, as it were, in the air.")

The fugals listed on the right side of the table are a bit less standard but represent a useful breakdown. Immobile factors—certainly land and natural resources, and in an international context people as well—militate against concentration of production, both from the supply side (some production must go to where the workers are) and from the demand side (dispersed factors create a dispersed market, and some production will have an incentive to locate close to the consumers). Concentrations of economic activity generate increased demand for local land, driving up land rents and thereby providing a disincentive for further concentration. And concentrations of activity can generate more or less pure external diseconomies such as congestion.

In the real world not only agglomeration in general, but any particular example of agglomeration, typically reflects *all* items on the menu. Why is the financial services industry concentrated in New York? Partly because the sheer size of New York makes it an attractive place to do business, and the concentration of the financial industry means that many clients and many ancillary services are located there. But a thick market for those with special skills, such as securities lawyers, and the general importance of being in the midst of the buzz are also important. Why doesn't all financial business concentrate in New York? Partly because many clients are *not* there, partly because renting office space in New York is expensive, and partly because dealing with the city's traffic, crime, and so on is such a nuisance.

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 column A and one from column B: 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 paper and book (Krugman 1991a, b) the normal choice has been the first item in each column: linkages as the force for concentration, immobile resources creating the tension necessary to keep the model interesting.

These choices are dictated less by empirical judgment than by two strategic modeling considerations. First, it is desirable to put some distance between the assumptions and the conclusions—to avoid something that looks too much like the assertion that agglomeration takes place because of agglomeration economies. This is especially true because much of the analysis we will 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 labelled "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. On the other hand, the thickness of the labor market, while it must have something to do with distance, does not lend itself quite so easily to being placed in a spatial setting. And land rents as a centrifugal force turn out to pose some conceptual issues—the "infinite Los Angeles problem"—which will be discussed briefly in the section on chance and necessity below.

In short, work to date on the "new economic geography" has been driven by considerations of modeling strategy to concentrate on the role of market-size effects in

generating linkages that foster geographical concentration, on one hand, and the opposing force of immobile factors working against such concentration, on the other.

Modeling Tricks

The idea is hardly a new one that there may be a circular process in which the decision of individual producers to choose a location with good access to markets and suppliers actually improves the market or supply access of other producers in that location. Indeed, it was the central theme of studies by Harris (1954) and Pred (1966) 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 at all: they would simply supply consumers from many local plants. And expansion of a regional market would not predictably lead to any increase in the range of goods produced within 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 there are important economies of scale enforcing 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 producing and delivering his product 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 there are economic activities that 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 that geography has finally made it into the economic mainstream is therefore obvious: imperfect competition is no longer regarded as impossible to model, and so stories that crucially involve unexhausted scale economies are no longer out of bounds. Indeed, the new interest in geography may be regarded as the fourth (and final?) wave of the increasing returns/imperfect competition revolution that has swept through economics over the past two decades. 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 perhaps be best described as a "genre," or style of economic analysis which tries to explain the spatial structure of the economy using certain technical tricks to produce models in which there are increasing returns and markets characterized by imperfect competition. These tricks are summarized by Fujita, Krugman, and Venables (forthcoming) with the slogan "Dixit-Stiglitz, icebergs, evolution, and the computer." Let us consider each part of that slogan in turn.

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 the modeler respect the integer nature of many location decisions—no fractional plants allowed—yet analyze his model in terms of the behavior of continuous variables like 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. This is a less familiar technical trick. In location theory, transportation costs are of the essence. Yet any attempt to develop a general-equilibrium model of economic geography would be substantially complicated by the need to model the transportation as well as the 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 Paul Samuelson (1952) in international trade theory: that a fraction of any good shipped simply "melts away" in transit, so that transport costs are in effect incurred in the good shipped. (In the new 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 modelling convenience there turns out to be a spectacular synergy between 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 f.o.b. (free on board) 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 move gradually toward locations that offer higher current real returns. This sort of dynamic process was initially proposed apologetically, since 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 like myself, it sometimes seems that the main contribution of evolutionary game theory has been to relegitimize those little arrows we always wanted to draw on our diagrams.)

THE COMPUTER. Finally, despite the best efforts of the theorist, all but the simplest models of economic geography usually turn out to be a bit 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 Geographical Change

Suppose that for some reason some 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 for some reason 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 will 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 models are characterized by bifurcations. And bifurcation diagrams are therefore a central analytical tool in this literature.

The typical form of these bifurcations may be illustrated by figures 1 and 2, which show results from a simulation of the model originally introduced in Krugman 1991a. That paper was, in effect, an attempt to formalize the story suggested by Harris and Pred. The model envisaged an economy consisting of two symmetric regions 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 the centrifugal force.

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

The rough intuition behind these curves runs as follows. In the case of high transport costs, there is relatively little interregional trade. So the wages workers can earn depend mainly on the amount of local competition and are thus decreasing in the number of other workers in the same region. On the other hand, when transport costs are low, a typical firm sells extensively in both regions. But since it has better access to markets if it is located in the region with the larger population, it can afford to pay higher wages—and the purchasing power of these wages is also higher because workers have better access to consumer goods. So in that case real wages are increasing in a region's population. At intermediate transport costs these two forces are nearly balanced. The particular curvature shown, in which centripetal forces are stronger when regions are very unequal, while centrifugal forces are stronger when they are nearly symmetric, is an artifact of the particular functional forms used in this exercise.

Since 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, 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 move gradually 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 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 the point labelled *A*. Then suppose that transport costs were gradually to fall. When the economy reached *B*, it would begin a cumulative process, in which a growing concentration of manufacturing in one region would lead to a still larger concentration of manufacturing in that region. That is, the economy would spontaneously organize itself into a core-periphery geography.

Geographical 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, though less advanced. Rather, they argued, the emergence of both rich and poor countries was part of a common process of uneven development, in which some initial advantages on the part of 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 last decade, of course, worries have largely reversed: now it is the advanced countries that seem to fear that the rise of newly industrializing economies will undermine their prosperity.

It turns out that new geography-type models can actually help shed light on both concerns. They indeed 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 as a result of 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 then 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), while 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 costs of transporting manufactured goods between them. If transport costs are high, each region will be essentially self-sufficient, and the regions will therefore be symmetric in outcomes as well as initial conditions. But now imagine gradually reducing 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 expenditure. If this share is low, the region that becomes the core does not get a significantly higher wage rate from that role. But if it is sufficiently large (in a two-region model, if it exceeds one-half of the total expenditure on traded goods), then 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). However, as the level of transport costs declines 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 paper as "history of the world, Part I."

I will come back shortly to the question of how much of the history of the real world such an analysis actually captures. First, however, it will be useful to use the geographical 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 the existence of national boundaries, asking how an undifferentiated, "seamless" world economy might evolve 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. However, it is possible to get considerable insight by focusing on particular, unrealistic, but convenient "geometries" for the world.

One particular geometry that is useful in spite of 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 computations. Second, since 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-Venables model (1997) 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 rather 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, which determines the "phase" of the regional development pattern (that is, 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. However, one can say that 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 certainly 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 within Developing Countries

It is often been observed both 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 typically has a strong geographic dimension. Although much 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 immobiles 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 paper, the original source of the south's advantage need not lie in any inherent superiority of its resources or location: it could be simply a result of historical accident.

While this is a coherent story, however, some modeling of regional inequalities has suggested an additional source of those inequalities: self-reinforcing advantages of market access via transportation networks.

One simple version of this story was laid out in Krugman (1993) and is illustrated in the left panel of figure 3. The figure shows three locations; the width of the lines between the locations is an inverse indicator of transportation costs (that is, thicker lines mean lower transport 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 1 to either of the other locations than it is to go

between those locations. What is easy to show is that other things being equal—that is, 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 one of those painfully obvious points that somehow just wasn't in the literature before.)

But why should transport costs be lower between 1 and other locations than between those other locations? One obvious answer is that if industry is concentrated in 1, there will be more trade between 2 and 1 than between 2 and 3, and so on. And if there are increasing returns in transportation—as there surely are—this will mean lower per-unit costs of transportation 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. In Krugman (1993) it is shown 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 1 and 2 than between either of those regions and the rest of the economy, and with transport costs between 3 and 4 particularly high. This pattern might emerge, again in the presence of increasing returns in transportation, if 1 and 2 both had large concentrations of industry. The effect, of course, would be to make 1 and 2 more attractive places to do business, reinforcing their advantage. A concrete

example: part of the advantage of São Paulo is its good access to Rio, including frequent plane flights, and vice versa. This is natural between the two largest cities of Brazil, but further reinforces the tendency of activity to concentrate in those two cities.

Just as in the global economy models discussed in section 2, models of regional inequality can easily show a nonmonotonic response to declining transport costs. Initially the effect of such declines can be to promote the formation of core-periphery patterns. To take a classic example, the stark division of Italy into affluent north and Mezzogiorno took shape when railroads were introduced. These railroads had the effect of making it possible for factories in the north to supply the needs of agricultural markets in the south, causing deindustrialization in the south. Moreover, the initial railroad net inevitably 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 at small scale of transportation) can lead to a spread of industry: once it is inexpensive to transport inputs to 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 of their levels in São Paulo. This is one of the factors often blamed for the rising unemployment in the 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 (no surprise) and, more interestingly, political structure: federal or decentralized systems do not have primate cities as large as those with high centralization. Thus Mexico City is still larger than Shanghai, because of China's decentralization.

The role of political centralization in causing primacy is at one level fairly obvious: it results both from the direct demand and employment created by the government apparatus and 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): whatever initial concentration of demand and advantages of access are conveyed to businesses in the capital will be magnified via the usual set of 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 (for example, the fact that Tokyo is

substantially larger than New York even though Japan has only half as many people as the United States).

There may also be other important policy linkages. Hanson (1992) pointed out 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 (1996) an effort was made to justify this observation in terms of a formal model. The paper 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 close to the large *domestic* market—and would also buy many of their inputs from abroad—and hence have less incentive to locate close to domestic suppliers. Meanwhile, high land rents would still create an incentive to locate away from other producers. Numerical examples confirmed that high trade barriers would tend to foster concentration of manufacturing in a single Mexico City–type location, while 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, since proximity to the border is not an issue—indeed, given Mercosur the border issue actually cuts the other way.)

For what it is worth, cross-sectional regressions by Ades and Glaser (1995) do in fact find some 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 in general difficult for that reason.)

Chance and Necessity

At the beginning of this paper 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 Jeffrey Sachs, 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 the specific example of 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 pre-industrial standards). It was quite natural that this location should have become the site for Mexico's major urban center. On the other hand, there is no longer a lake, or for that matter any agriculture to speak of in the valley. 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 prime city was dictated by natural geography; yet those geographical 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 its dominant position is a classic first-world example of the same proposition.)

Or to use a somewhat different metaphor: in many cases, aspects of natural geography are able to matter so much not because natural features of the landscape are that crucial, but because they establish seeds around which self-reinforcing agglomerations crystallize. 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 quite 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 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 such small inherent differences strongly bias the outcome when one starts with some sort of random allocation of mobile factors.)

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

We 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 gradually increasing the population, eventually it becomes profitable for some manufacturing to locate away from the original center, and new cities emerge.

But where do they emerge? Figure 4 represents a version of Fujita and Mori's analysis. We imagine that the economy is in a long, narrow valley (making it effectively one-dimensional), with the original city at A. We put a fork in the valley at B, so that B 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.

Now as the population expands, we will see A's agricultural hinterland expand as well, eventually pushing up both forks of the valley beyond B. And eventually also a new city will emerge. Where? The answer is that B is an "especially likely" location, in the following sense: imagine choosing alternative initial positions for A (or varying any other parameter of the model)

and asking where the next city emerges. In general, any possible location will be chosen for at most one location of the original city. However, because of the special advantages of 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 at B. So the natural geography will often though 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 the existence of strong circular causation has some important implications for the interpretation of correlations found between natural advantages and actual economic geography: they 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 seeds around which cities crystallize explains why most large cities today are ports. However, because the importance of the port was *only* that of serving as a seed, 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 well become self-sustaining even though its location does not fit any of the criteria that characterize major existing cities.

To put a sharper point on it: the current pattern of world economic geography clearly shows a strong association between per capita income and more or less Western European conditions—temperate climate, absence of malaria, much of the population close to the coast or navigable rivers or both. But this may mainly reflect the catalytic role of these factors in the past

and need not say that an inland country (which now has access to good roads and cheap air transport) with a hot climate (but now with 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 which brings us to policy.

Geography and Policy

This will, necessarily, be a short concluding section. At this point there has been quite little effort to draw policy conclusions from the "new economic geography" literature. The main thing for the moment is to explain why.

In principle, the sort of economy envisaged by the models sketched out in this paper ought to be a prime target for government intervention. Clearly there is no presumption here that the market gets it right. Moreover, the models suggest that under some circumstances small policy interventions can have large and perhaps lasting effects. Finally, since the cumulative processes of concentration tend to produce winners and losers, perhaps at the level of nations, there is an obvious incentive to try to make sure that your own 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 if anything new geography models—in which the crucial effects are general-equilibrium rather than merely partial-equilibrium—are if anything likely to be even harder to make operational.

There is also, 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 more or less frantic efforts of interested parties to recruit some 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 is also a special consideration that makes policy conclusions difficult in the geographical literature. Consider table 1 again, bearing in mind that normally *all* of 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 we would generate by having more activity here). One may have opinions—I am quite sure in my gut, and even more so my lungs, that Mexico City is too big—but gut feelings are not a sound basis for policy.

Fortunately, we can make a safe recommendation: since 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: Fischer Verlag.

Davis, D., and R. Weinstein. 1997. “Empirical Testing of Economic Geography: Evidence from Regional Data.” Harvard University, Cambridge, Mass.

Dicken, P., and P. Lloyd. 1990. *Location in Space: Theoretical Perspectives in Economic Geography*. New York: HarperCollins.

Dixit, A., and J. Stiglitz. 1977. “Monopolistic Competition and Optimum Product Diversity.” *American Economic Review* 67: 297–308.

Fujita, M., and P. Krugman. 1995. “When is the Economy Monocentric: von Thünen and Christaller Unified.” *Regional Studies and Urban Economics* 25: 505–28.

Fujita, M., P. Krugman, and A. Venables. Forthcoming. *The Spatial Economy*.

Fujita, M., and T. Mori. 1996a. “Structural Stability and Evolution of Urban Systems.” *Regional Science and Urban Economics.*

———. 1996b. “The Role of Ports in the Making of Major Cities: Self-Agglomeration and Hub-Effect.” *Journal of Development Economics* 49: 93–120.

Fujita, M., T. Mori, and P. Krugman. 1997. “On the Evolution of Hierarchical Urban Systems.” Kyoto.

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, E., and R. Vernon. 1959. *Anatomy of a Metropolis.* Cambridge, Mass.: Harvard University Press.

Karaska, G., and D. Bramhall, eds. 1969. *Locational Analysis for Manufacturing.* Cambridge, Mass.: MIT Press.

Krugman, P. 1991a. "Increasing Returns and Economic Geography." *Journal of Political Economy.*

———. 1991b. *Geography and Trade.* Cambridge, Mass.: MIT Press.

———. 1993. "On the Number and Location of Cities." *European Economic Review.*

Krugman, P., and R. Livas Elizondo. 1996. "Trade Policy and the Third World Metropolis." *Journal of Development Economics* 49: 137–50.

Krugman, P., and A. 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." Massachusetts Institute of Technology, Cambridge, Mass.

Lösch, A. 1940. *The Economics of Location.* Jena: Fischer Verlag.

Pred, A. R. 1966. *The Spatial Dynamics of U.S. Urban-Industrial Growth, 1800–1914.*

Puga, D., and A. Venables. 1997. "The Spread of Industry: Spatial Agglomeration in Economic Development." CEPR Working Paper 1354. Centre for Economic Policy Research, London.

Samuelson, P. 1954. "The Transfer Problem and Transport Costs." *Economic Journal* 64: 264–89.

Turing, A. 1952. “The Chemical Basis of Morphogenesis.” *Philosophical Transactions of the Royal Society of London* 237: 37.

Venables, A. 1996. “Equilibrium Locations of Vertically Linked Industries.” *International Economic Review.*

von Thünen, J. 1826. *The Isolated State.* London: Pergamon.

Weber, A. 1909. *Theory of the Location of Industries.* Chicago: University of Chicago Press.

Weibull, M. 1995. *Evolutionary Game Theory.* Cambridge, Mass.: MIT Press.