NBER WORKING PAPER SERIES

INCREASING RETURNS AND ECONOMIC GEOGRAPHY

Paul Krugman

Working Paper No. 3275

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 1990

NBER Working Paper #3275
March 1990

INCREASING RETURNS AND ECONOMIC GEOGRAPHY

ABSTRACT

This paper develops a two-region, two-sector general equilibrium model of location. The location of agricultural production is fixed, but monopolistically competitive manufacturing firms choose their location to maximize profits. If transportation costs are high, returns to scale weak, and the share of spending on manufactured goods low, the incentive to produce close to the market leads to an equal division of manufacturing between the regions. With lower transport costs, stronger scale economies, or a higher manufacturing share, circular causation sets in: the more manufacturing is located in one region, the larger that region's share of demand, and this provides an incentive to locate still more manufacturing there. Thus when the parameters of the economy lie even slightly on one side of a critical "phase boundary", all manufacturing production ends up concentrated in only one region.

Paul Krugman
Department of Economics
MIT
50 Memorial Drive
Cambridge, MA 02139
(617)253-2655

The study of economic geography -- of the location of factors of production in space -- occupies a relatively small part of standard economic analysis. International trade theory, in particular, conventionally treats nations as dimensionless points (and frequently assumes zero transportation costs between countries as well). Admittedly, models descended from von Thünen play an important role in urban studies, while Hotelling-type models of locational competition get a reasonable degree of attention in industrial organization. On the whole, however, it seems fair to say that the study of economic geography plays at best a marginal role in economic theory.

On the face of it, this neglect is surprising. The facts of economic geography are surely among the most striking features of real-world economies, at least to laymen. For example, one of the most remarkable things about the United States is that in a generally sparsely populated country, much of whose land is fertile, the bulk of the population resides in a few clusters of metropolitan areas; forty percent are crowded into a not especially inviting section of the East Coast. It has often been noted that nighttime satellite photos of Europe reveal little of political boundaries, but clearly suggest a center-periphery pattern whose hub is somewhere in or near Belgium. A layman might have expected that these facts would play a key role in economic modelling.

Furthermore, there is a long if somewhat thin tradition in location theory, that one might have supposed would inspire the efforts of both theorists and econometricians. Indeed, several schools of thought may be identified. Best known, perhaps, is the German School, originating in the work of von Thünen (1826) but led in the twentieth century by Weber (1909), Christaller (1933), and Lösch (1940). Inspired by this German school, but less preoccupied with the geometry of location, was the American school of regional science, including Hoover (1948) and especially Isard (1956). Yet another

tradition, drawing on Marshall's initial description of agglomeration due to external economies, stresses the role of externalities in producing divergent regional development; the most influential writings in this tradition are those of Myrdal (1957), Hirschmann (1958), and Perroux (1950), and this tradition has been carried on more recently by David (1984) and Arthur (1989).

Economic geography, then, is both an important subject and one that has at least occasionally drawn sustained attention. Yet it is largely ignored by the economics profession. Why? The answer seems clearly to lie in considerations of method. The interesting questions of economic geography are not easily addressed by the model of competitive general equilibrium that increasingly came to dominate economic thinking between 1940 and the 1970s. If we ask why so much of the American economy is concentrated in a few coastal strips, we are immediately driven to speak about economies of scale and externalities. Yet economies of scale internal to firms imply imperfect competition, which until recently was regarded as too difficult to model rigorously, while purely technological external economies seem both implausible and too elusive to have useful empirical content. The result is that discussions of economic geography have historically tended to rely either on logically incomplete models or on verbal discussion in which models are at best implicit. As standards of rigor rose over time, and as those economists who wrote about geographical issues failed to keep up, their work and the subject as a whole was simply submerged.

This crowding out of important but poorly formalized insights in economic geography is reminiscent of what happened in several other areas of economics. Most notably, in international trade the insights of such thinkers as Burenstam Linder (1961), Vernon (1966), and even of important parts of Ohlin were increasingly neglected as a rigorous general equilibrium approach became de rigeur; while in development economics the same happened to the ideas of

such authors as Young (1928) and Rosenstein-Rodan (1943). In both trade and development, however, recent applications of new models derived from industrial organization have begun to restore the prominence of these earlier ideas. In trade, the "new international economics" of such authors as Dixit and Norman (1980), Krugman (1979), and Helpman (1981) has given rigor, and hence respectability, to non-comparative-advantage explanations of trade. In development economics, the "new growth theory" of Romer (1986, 1987), Lucas (1988), and of Murphy, Shleifer, and Vishny (1989a, 1989b) has begun to accomplish the same thing.

The purpose of this paper is to suggest that application of models and techniques derived from theoretical industrial organization now allow a reconsideration of economic geography; that it is now time to attempt to incorporate the insights of the long but informal tradition in this area into formal models. In order to make the point, the paper develops a simple illustrative model designed to shed light on one of the key questions of location: why and when does manufacturing become concentrated in a few regions, leaving others relatively undeveloped?

What we will see is that it is possible to develop a very simple model of regional divergence based on the interaction of economies of scale with transportation costs. This is perhaps not too surprising, given the kinds of results that have been emerging in recent literature (with Murphy et. al. perhaps the closest parallel). More interesting is the fact that regional divergence need not always happen, and that whether it does depends in an interesting way on a few key parameters. These parameters define a sort of "phase boundary"; the geography of economies that lie on one side of that boundary will evolve in a fundamentally different way from that of economies that lie on the other side.

The paper is in five parts. The first part sets the stage with an

informal discussion of the problem. The second then sets out the analytical model. In the third part we analyze the determination of short-run equilibrium and dynamics. The fourth section analyzes the conditions under which regional divergence does and does not occur. Finally, the paper concludes with a brief discussion of some natural extensions.

1. Bases for regional divergence

Why is so much of the population of the US concentrated along 500 miles of the East Coast? The standard answer is "external economies". At some level this must be right; yet as it stands the answer is unsatisfying, because it is too vague and leaves too many questions hanging. What is the nature of these externalities? How necessary is the geographical concentration to their realization? How different would either history or technology have to have been for the great American megalopolis either not to exist or to be located somewhere else? Without more specificity these questions cannot usefully be posed.

I will adopt the working assumption that the externalities that sometimes lead to regional divergence are pecuniary externalities associated with either demand or supply linkages, rather than purely technological spillovers. In competitive general equilibrium, of course, pecuniary externalities have no welfare significance and could not lead to the kind of interesting dynamics we will derive later. Over the past decade, however, it has become a familiar point that in the presence of imperfect competition and increasing returns, pecuniary externalities matter -- for example, that if one firm's actions affect the demand for the product of another firm whose price exceeds marginal cost, this is as much a "real" externality as if one firm's R&D spills over

into the general knowledge pool. At the same time, by focussing on pecuniary externalities we are able to make the analysis much more concrete than if we allowed external economies to arise in some invisible form (this is particularly true when location is at issue: how far does a technological spillover spill?).

To understand the nature of the postulated pecuniary externalities, it is useful to retrace some of the steps of the grand tradition in location theory. Weber (1909), though best known for his "location triangles", also laid out a general view of the evolution of a pattern of location in a nation. He thought of this as involving the sequential laying down of a series of "strata", increasingly divorced from the distribution of natural resources. The first stratum would consist of farmers, miners, etc.. whose location would be determined by the distribution of arable land and other resources. One might idealize the distibution of this first stratum by imagining it uniformly spread across a featureless plain. The second stratum would consist of less locationally bound activities designed to service the first stratum -- market towns, manufacturing activities, and so on. Because of transportation costs, the second stratum's location would tend to follow that of the first; because of economies of scale, however, it would not be uniformly distributed. Instead, it would form a sort of lattice across the plain. There would then be a third stratum of activities servicing the second stratum, and forming a sparser lattice, and so on.

Two later authors elaborated these basic scheme. Christaller (1933) argued that the lattices of the second, third, etc. strata would form a hierarchy of central places, whose number would decrease but population increase as one went up the scale. Christaller documented the existence of such a hierarchy in southern Germany. Lösch (1940), in a famous contribution, pointed out that if the objective was to minimize transportation costs, then

the lattice of central places on a featureless plain would form a series of hexagonal market areas.

But is this scheme right? Isard (1956) pointed out a key problem with Weber's view, and hence with the Christaller-Lösch extensions. According to the Weberian story, the second stratum exists to service the first, the third to service the second, and so on. However, some of the demand for the second stratum's services will come, not from the first stratum, but from the second and higher strata themselves. This immediately raises the possibility of a process of circular causation: the location of higher strata depends on the distribution of demand, but the distribution of demand depends on the location of higher strata.

The circularity become still worse if one takes into account another factor: it will, other things equal, be more desirable to live and produce near a central place high in the hierarchy, because it will than be less expensive to buy the goods and services this central place provides.

The circularity will not matter too much if the higher strata employ only a small fraction of the population and hence generate only a small fraction of demand; or if a combination of weak economies of scale and high transportation costs induce suppliers of goods and services to the lowest stratum to locate very close to their markets. These criteria would have been satisfied in a pre-railroad, pre-industrial society, such as that of sixteenth-century Europe. In such a society the bulk of the population would have been engaged in agriculture; the small manufacturing and commercial sector would not have been marked by very substantial economies of scale; and the costs of transportation would have ensured that most of the needs that could not be satisfied by rural production would be satisfied by small towns serving roughly hexagonal market areas.

But now let the society become richer, so that a higher fraction of

income is spent on non-agricultural goods and services; let the factory system and eventually mass production emerge, and with them economies of large-scale production; and let canals, railroads, and finally automobiles lower transportation costs. Then the tie of production to the distribution of land will be broken. A region with a relatively large nonrural population will be an attractive place to produce both because of the large local market and because of the availability of the goods and services produced there; this will attract still more population, at the expense of regions with smaller initial production; and the process will feed on itself until the whole of the nonrural population is concentrated in a few regions.

There are two interesting points suggested by this imaginary history. First, it seems that small changes in the parameters of the economy may have large effects on its qualitative behavior. That is, when some index that takes into account transportation costs, economies of scale, and the share of nonagricultural goods in expenditure crosses a critical threshhold, population will start to concentrate and regions to diverge; once started. this process will feed on itself. Thus the geography will go through a kind of change of state when the index crosses a critical level, much as water changes its qualitative behavior when the temperature goes from a little above to a little below freezing.

Second, the details of the geography that emerges -- which regions end up with the population -- depend sensitively on initial conditions. If one region has slightly more population than another when, say, transportation costs fall below some critical level, that region ends up gaining population at the other's expense; had the distribution of population at that critical moment been only slightly different, the roles of the regions might have been reversed. Again to use a physical analogy, this is a "random broken symmetry": like ice crystallizing as water is cooled, the detailed structure depends on

possibly small accidents of early history.

This is about as far as an informal story can take us. The next step is to develop as simple as possible a formal model, to see whether the story just told can be given a more rigorous formulation.

2. A two-region model

We begin, for simplicity, with a model of two regions (a many-region model is considered in the last section). In this model there are assumed to be two kinds of production: agriculture, which is a constant-returns sector tied to the land, and manufactures, an increasing-returns sector that can be located in either region.

The model, like many of the models in both the new trade and the new growth literature, is a variant on the monopolistic competition framework initially proposed by Dixit and Stiglitz (1977). This framework, while admittedly special, is remarkably powerful in its ability to yield simple intuition-building treatments of seemingly intractable issues.

All individuals in this economy, then, are assumed to share a utility function of the form

$$U = C_M^{\mu} C_A^{1-\mu} \tag{1}$$

where C_A is consumption of the agricultural good and C_M is consumption of a manufactures aggregate. Given (1), of course, manufactures will always receive a share μ of expenditure; this share is one of the key parameters that will determine whether regions converge or diverge.

The manufactures aggregate C_M is defined by

$$C_M = \left(\sum_{i=1}^{N} c_i^{(\sigma-1)/\sigma}\right)^{\sigma/(\sigma-1)} \qquad (2)$$

where N is the large number of potential products and $\sigma>1$ is the elasticity of substitution among the products. The elasticity σ is the second parameter determining the character of equilibrium in the model.

There are two regions in the economy, and two factors of production in each region. Following the simplification suggested in Krugman (1981), each factor is assumed specific to one sector. Peasants produce agricultural goods; without loss of generality we suppose that the unit labor requirement is one. The peasant population is assumed completely immobile between regions, with a given peasant supply $(1-\mu)/2$ in each region. Workers may move between the regions; we let L_1, L_2 be the worker supply in regions 1 and 2, respectively, and require only that the total add up to the overall number of workers μ[1]:

$$L_1 + L_2 = \mu \qquad (3)$$

The production of an individual manufactured good i involves a fixed cost and constant marginal cost, giving rise to economies of scale:

$$\mu_i = \alpha + \beta x_i \qquad (4)$$

where μ_i is the labor used in producing i and x_i is the good's output.

We turn next to the structure of transportation costs between the two regions. Two strong assumptions will be made for tractability. First,

[1]This choice of units ensures that the wage rate of workers equals that of peasants in long-run equilibrium.

transportation of agricultural output will be assumed to be costless.[2] The effect of this assumption is to ensure that the price of agricultural output, and hence the earnings of each peasant, are the same in both regions. We will use this common agricultural price/wage rate as numeraire. Second, transportation costs for manufactured goods will be assumed to take Samuelson's "iceberg" form, in which transport costs are incurred in the good transported. Specifically, of each unit of manufactures shipped from one region to the other, only a fraction $\tau<1$ arrives. This fraction τ, which is an inverse index of transportation costs, is the final parameter determining whether regions converge or diverge.

We can now turn to the behavior of firms. Suppose that there are a large number of manufacturing firms, each producing a single product. Then given the definition of the manufacturing aggregate (2) and the assumption of iceberg transport costs, the elasticity of demand facing any individual firm is σ (see Krugman (1980)). The profit-maximizing pricing behavior of a representative firm in region 1 is therefore to set a price equal to

$$p_1 = [\sigma/(\sigma-1)]\beta w_1 \tag{5}$$

where w_1 is the wage rate of workers in region 1. A similar equation applies in region 2; so comparing the prices of representative products, we have

[2]The reason for this assumption is the following: since agricultural products are assumed to be homogeneous, each region is either exporting or importing them, never both. But if agricultural goods are costly to transport, this would introduce a "cliff" at the point where the two regions have equal numbers of workers, and thus where neither had to import food. This is evidently an artifact of the two-region case: if peasants were spread uniformly across a featureless plain, there would be no discontinuity.

$$p_1/p_2 = w_1/w_2 \tag{6}$$

If there is free entry of firms into manufacturing, profits must be driven to zero. Thus it must be true that

$$(p_1-\beta w_1)x_1 = \alpha w_1 \tag{7}$$

which implies

$$x_1 = x_2 = \alpha/(\sigma-1) \tag{8}$$

That is, output per firm is the same in each region, irrespective of wage rates, relative demand, etc.. This has the useful implication that the number of manufactured goods produced in each region is proportional to the number of workers, so that

$$n_1/n_2 = L_1/L_2 \tag{9}$$

It should be noted that in zero-profit equilibrium $\sigma/(\sigma-1)$ is the ratio of the marginal product of labor to its average product, i.e., the degree of economies of scale. Thus σ, although it is a parameter of tastes rather than technology, can be interepreted as an inverse index of equilibrium economies of scale. As such, it is the third and final parameter determining the behavior of this economy.

We have now laid out the basic structure of the model. The next step is to turn to the determination of equilibrium.

3. Short-run and long-run equilibrium

This model lacks any explicit dynamics. However, it is useful to have a concept of short-run equilibrium before turning to full equilibrium. Short-run equilibrium will be defined in a Marshallian way, as an equilibrium in which the allocation of workers between regions may be taken as given. We then suppose that workers move toward the region that offers them higher real wages, leading either to convergence between regions as they move toward equality of worker-peasant ratios, or divergence as the workers all congregate in one region.

To analyze short-run equilibrium, we begin by looking at the demand within each region for products of the two regions. Let c_{11} be the consumption in region 1 of a representative region 1 product, and c_{12} be the consumption in region 1 of a representative region 2 product. The price of a local product is simply its f.o.b. price p_1; the price of a product from the other region, however, is its transport-cost-inclusive price p_2/τ. Thus the relative demand for representative products is

$$c_{11}/c_{12} = (p_1\tau/p_2)^{-\sigma} = (w_1\tau/w_2)^{-\sigma} \tag{10}$$

Define z_{11} as the ratio of region 1 expenditure on local manufactures to that on manufactures from the other region. Two points should be noted about z. First, a one percent rise in the relative price of region 1 goods, while reducing the relative quantity sold by σ percent, will reduce the value by only σ-1 percent, because of the valuation effect. Second, the more goods produced in region 1, the higher their share of expenditure for any given relative price. Thus z_{11} equals

$$z_{11} = (n_1/n_2)(p_1^r/p_2)(c_{11}/c_{12}) = (L_1/L_2)(w_1^r/w_2)^{-(\sigma-1)} \quad (11)$$

Similarly, the ratio of region 2 spending on region 1 products to spending on local products is

$$z_{12} = (L_1/L_2)(w_1/w_2^r)^{-(\sigma-1)} \quad (12)$$

The total income of region 1 workers is equal to the total spending on these products in both regions. (Transportation costs are included because they are assumed to be incurred in the goods themselves). Let Y_1, Y_2 be the regional incomes (including the wages of peasants). Then the income of region 1 workers is

$$w_1L_1 = \mu\{[z_{11}/(1+z_{11})]Y_1 + [z_{12}/(1+z_{12})]Y_2\} \quad (13)$$

and the income of region 2 workers is

$$w_2L_2 = \mu\{[1/(1+z_{11})]Y_1 + [1/(1+z_{12})Y_2\} \quad (14)$$

The incomes of the two regions, however, depend on the distribution of workers and their wages. Recalling that the wage rate of peasants is the numeraire, we have

$$Y_1 = (1-\mu)/2 + w_1L_1 \quad (15)$$

$$Y_2 = (1-\mu)/2 + w_2L_2 \quad (16)$$

The set of equations (11)-(16) may be regarded as a system that

determines w_1 and w_2 (as well as four other variables) given the distribution of labor between regions 1 and 2. By inspection, one can see that if $L_1=L_2$, $w_1=w_2$. If labor is then shifted to region 1, however, the relative wage rate w_1/w_2 can move either way. The reason is that there are two opposing effects. On one side, there is the "home market effect": other things equal, the wage rate will tend to be higher in the larger market (see Krugman (1980)). On the other side, there is the extent of competition: workers in the region with the smaller manufacturing labor force will face less competition for the local peasant market than those in the more populous region. In other words, there is a tradeoff between proximity to the larger market and lack of competition for the local market.

In moving from short-run to long-run equilibrium, however, a third consideration enters the picture. Workers are interested, not in nominal wages, but in real wages; and workers in the region with the larger population will face a lower price for manufactured goods. Let $f=L_1/\mu$, the share of the manufacturing labor force in region 1. Then the true price index of manufactured goods for consumers residing in region 1 is

$$P_1 = [fw_1^{-(\sigma-1)} + (1-f)(w_2/\tau)^{-(\sigma-1)}]^{-1/(\sigma-1)} \qquad (17)$$

while that for consumers residing in region 2 is

$$P_2 = [f(w_1/\tau)^{-(\sigma-1)} + (1-f)w_2^{-(\sigma-1)}]^{-1/(\sigma-1)} \qquad (18)$$

and the real wages of workers in each region are

$$\omega_1 = w_1 P_1^{-\mu} \qquad (19)$$

$$\omega_2 = w_2 P_2^{-\mu} \tag{20}$$

Looking at (17) and (18), it is apparent that if wage rates in the two regions are equal, a shift of workers from region 2 to region 1 will lower the price index in region 1 and raise it in region 2, and thus raise real wages in region 1 relative to those in region 2. This therefore adds an additional reason for divergence.

We may now ask the crucial question: how does ω_1/ω_2 vary with f? We know by symmetry that when f=1/2, i.e., when the two regions have equal numbers of workers, they offer equal real wage rates. But is this a stable equilibrium? It will be if ω_1/ω_2 decreases with f; for in that case whenever one region has a larger work force than the other, workers will tend to migrate out of that region. In this case we will get regional convergence. On the other hand, if ω_1/ω_2 <u>increases</u> with f, workers will tend to migrate <u>into</u> the region that already has more workers, and we will get regional divergence.[3] As we have seen, there are two forces working toward divergence -- the home market effect and the price index effect -- and one working toward convergence, the degree of competition for the local peasant's market. The question is which forces dominate.

In principle, it is possible simply to solve our model for real wages as a function of f. This is, however, difficult to do analytically. In the next section an indirect approach is used to characterize the model's behavior. For now, however, let us simply note that there are only three parameters in this

[3]Strictly speaking, a dynamic story should take expectations into account. It is possible that workers may migrate into the region that initially has fewer workers, because they expect other workers to do the same. This kind of self-fulfilling prophecy can only occur, however, if adjustment is rapid and discount rates not too high. See Krugman (1989) for an analysis.

model that cannot be eliminated by choice of units: the share of expenditure on manufactured goods, μ; the elasticity of substitution among products, σ; and the fraction of a good shipped that arrives, τ. And the model can be quite easily solved numerically for a variety of parameters. Thus it is straightforward to show that depending on the parameter values we may have either regional convergence or regional divergence.

Figure 1 makes the point. It shows computed values of ω_1/ω_2 as a function of f in two different cases. In both cases we assume σ=4 and μ=.3. In one case, however, τ=.5 (high transportation costs), while in the other τ=.75 (low transportation costs). In the high transport cost case, the relative real wage declines as f rises. Thus in this case we would expect to see regional convergence, with the geographical distribution of the "second stratum" following that of the first. In the low transport cost case, however, the slope is reversed; thus we would expect to see regional divergence.

It is possible to proceed entirely numerically from this point, generating a "map" of parameter values for which convergence or divergence will occur. By taking a somewhat different approach, however, it is possible to characterize the properties of this map analytically, and also to develop a simple way of computing it.

4. Convergence vs. divergence

To ask when regions diverge, it turns out to be most useful to reverse the way we approach the problem. Instead of asking whether an equilibrium in which workers are distributed equally between the regions is stable, we ask whether a situation in which all workers are concentrated in one region is an equilibrium.

Consider a situation in which all workers are concentrated in region 1

(the choice of region of course is arbitrary). Region 1 will then constitute a larger market than region 2. Since a share of total income μ is spent on manufactures, and all of this income goes to region 1, we have

$$Y_2/Y_1 = (1-\mu)/(1+\mu) \tag{21}$$

Let n be the total number of manufacturing firms; then each firm will have a value of sales equal to

$$V_1 = (1/n)(Y_1 + Y_2) \tag{22}$$

which is just enough to allow each firm to make zero profits.

Now we ask: is it possible for an individual firm to commence production profitably in region 2? (I will refer to such a hypothetical firm as a "defecting" firm). If not, then concentration of production in region 1 is an equilibrium; if so, it isn't.

In order to produce in region 2, a firm must be able to attract workers. To do so, it must compensate them for the fact that all manufactures (except its own infinitesimal contribution) must be imported; thus we must have

$$w_2/w_1 = (1/\tau)^{\mu} \tag{23}$$

Given this higher wage, the firm will charge a profit-maximizing price that is higher than that of other firms in the same proportion. We can use this fact to derive the value of the firm's sales. In region 1, the defecting firm's value of sales will be the value of sales of a representative firm times $(w_2/w_1\tau)^{-(\sigma-1)}$. In region 2, its value of sales will be that of a

representative firm times $(w_2\tau/w_1)^{-(\sigma-1)}$. So the total value of the defecting firm's sales will be

$$V_2 = (1/n)[(w_2/w_1\tau)^{-(\sigma-1)}Y_1 + (w_2\tau/w_1)^{-(\sigma-1)}Y_2] \tag{24}$$

Notice that transportation costs work to the firm's disadvantage in its sales to region 1 consumers, but work to its advantage on sales to region 2 consumers (because other firms must pay them but it does not).

From (22), (23), and (24) we can (after some manipulation) derive the ratio of the value of sales by this defecting firm to the sales of firms in region 1:

$$V_2/V_1 = (1/2)\tau^{\mu(\sigma-1)}[(1+\mu)\tau^{(\sigma-1)}+(1-\mu)\tau^{-(\sigma-1)}] \tag{25}$$

One might think that it is profitable for a firm to defect as long as $V_2/V_1>1$, since firms will collect a constant fraction of any sales as a markup over marginal costs. This is not quite right, however, because fixed costs are also higher in region 2 because of the higher wage rate. So we must have $V_2/V_1>w_2/w_1 = \tau^{-\mu}$. We must therefore define a new variable,

$$\nu = (1/2)\tau^{\mu\sigma}[(1+\mu)\tau^{(\sigma-1)}+(1-\mu)\tau^{-(\sigma-1)}] \tag{26}$$

When $\nu<1$, it is unprofitable for a firm to begin production in region 2 if all other manufacturing production is concentrated in region 1. Thus in this case regional divergence is the long-run equilibrium. If $\nu>1$ concentration of production in one region is not an equilibrium.

Equation (26) at first appears to be a fairly unpromising subject for analytical results. However, it yields to careful analysis.

First note what we want to do with (26). It defines a boundary: a set of critical parameter values that mark the division between convergence and divergence. So we need only evaluate it in the vicinity of $\nu = 1$, asking how each of the three parameters must change in order to offset a change in either of the others.

Let us begin, then, with the most straightforward of the parameters, μ. We find that

$$\partial\nu/\partial\mu = \nu\sigma(\ln \tau) + (1/2)\tau^{\sigma\mu}[\tau^{\sigma-1} - \tau^{-(\sigma-1)}] < 0 \qquad (27)$$

That is, the larger the share of income spent on manufactured goods, the lower the relative sales of the defecting firm. This takes place for two reasons. First, workers demand a larger wage premium in order to move to the second region; this effect is reflected in the first term. Second, the larger the share of expenditure on manufactures, the larger the relative size of the region 1 market and hence the stronger the home market effect. This is reflected in the second term in (27).

Next we turn to transportation costs. From inspection of (26), we first note that when $\tau=1$, $\nu=1$ -- that is, when transport costs are zero location is irrelevant (no surprise!). Second, we note that when τ is small, ν approaches $(1-\mu)\tau^{1-\sigma(1-\mu)}$. Unless σ is very small or μ very large, this must exceed 1 for sufficiently small τ (the economics of the alternative case will be apparent shortly). Finally, we evaluate $\partial\nu/\partial\tau$:

$$\partial\nu/\partial\tau = \mu\sigma\nu/\tau + \tau^{\mu\sigma}(\sigma-1)[(1+\mu)\tau^{\sigma-1} - (1-\mu)\tau^{-(\sigma-1)}]/2\tau \qquad (28)$$

For τ close to 1, the second term in (28) drops out, leaving only the positive first term.

Taken together, these observations indicate a shape for ν as a function of τ that looks like Figure 2 (which represents an actual calculation for $\mu = .3$, $\sigma = 4$): at low levels of τ (i.e., high transportation costs), ν exceeds one and it is profitable to defect; at some critical value of τ, ν falls below one and concentrated manfacturing is an equilibrium, and the relative value of sales then approaches 1 from below.

The important point from this picture is that at the critical value of τ that corresponds to the boundary between convergence and divergence, $\partial\nu/\partial\tau$ is negative. That is, higher transportation costs militate against regional divergence.

We can also now interpret the case where $\sigma(1-\mu)<1$, so that $\nu<1$ even at arbitrarily low τ. This is a case where economies of scale are so large (small σ) and or the share of manufacturing in expenditure so high (high μ) that it is unprofitable to start a firm in region 2 no matter how high transport costs are.

Finally, we calculate $\partial\nu/\partial\sigma$. This equals

$$\partial\nu/\partial\sigma = \ln(\tau)\{\mu\nu + (1/2)\tau^{\mu\sigma}[(1+\mu)\tau^{\sigma-1} - (1-\mu)\tau^{-(\sigma-1)}]\}$$

$$= \ln(\tau)[\tau/\sigma](\partial\nu/\partial\tau) \qquad (29)$$

Since we have just seen that $\partial\nu/\partial\tau$ is negative at the relevant point, this implies that $\partial\nu/\partial\sigma$ is positive. That is, a higher elasticity of substitution (which also implies smaller economies of scale in equilibrium) works against regional divergence.

The implications of these results can be seen digrammatically. Holding σ constant, we can draw a "phase boundary" in μ, τ space. This boundary marks

parameter values at which firms are just indifferent between staying in a region-1 concentration or defecting. An economy that lies inside this boundary will not develop concentrations of industry in one or the other region, while an economy that lies outside the boundary will. The slope of the boundary is

$$\partial\tau/\partial\mu = -(\partial\nu/\partial\mu)/(\partial\nu/\partial\tau) < 0$$

If we instead hold μ constant and consider changing σ, we find

$$\partial\tau/\partial\sigma = -(\partial\nu/\partial\sigma)/(\partial\nu/\partial\tau) > 0$$

Thus an increase in σ will shift the boundary in μ,τ space outward.

Figure 3 shows calculated boundaries in μ,τ space for two values of σ, 4 and 10. The figure tells a simple story that is precisely the intuitive story given in part 1 of this paper. In an economy characterized by high transportation costs, a small share of footloose manufacturing, and/or weak economies of scale, the distribution of manufacturing production will be determined by the distribution of the "primary stratum" of peasants. With lower transportation costs, a higher manufacturing share, and/or stronger economies of scale, circular causation sets in, and manufacturing will concentrate in whichever region gets a head start.

What is particularly nice about this result is that it requires no appeal to elusive concepts like pure technological externalities; the external economies are pecuniary, arising from the desirability of selling to and buying from a region in which other producers are concentrated. It also involves no arbitrary assumptions about the geographical extent of external economies: distance enters naturally via transportation costs, and in no other way. The behavior of the model depends on "observable" features of the tastes

of individuals and the technology of firms; the interesting dynamics arise from interaction effects.

Obviously this is a highly special model. I will not attempt to generalize it significantly in this paper, but there is one special feature that needs some further discussion: the assumption that there are only two regions.

5. Multiple regions

The assumption of a two-region economy, while a natural first cut at this problem, begs many of the important questions of traditional economic geography. Among the extensions one should clearly try to make is therefore an effort to model multiple-region behavior.

In the grand tradition of economic geography one clearly ought to drop the notion of regions altogether, and start from a uniform distribution of peasants across a featureless two-dimensional plain. I would argue, however, that a premature attempt at quasi-realistic geometry has been one of the vices of economic geography, focussing its attention away from the fundamental economic issues. Preliminary insights can be gained without going this far. Specifically, let us assume that there are several distinct regions, each of which will itself be modelled as a point, and that the world is one-dimensional, i.e., the regions are laid out in a line.

It will be desirable to maintain symmetry; the only way to do this is to assume that the regions are in fact laid out in a circle (surrounding an impassable mountain range?). We also want to consider as few regions as possible consistent with interesting behavior; for reasons that will become apparent in a moment, the useful number turns out to be six.

Consider, then, the economy shown in Figure 4. It contains six regions, laid out in a circle. It has the same tastes and technology as our two-region model. Each region has one-sixth of the economy's peasant population. Manufactures production can be carried out in any region. However, when manufactures are shipped from one region to the next, a fraction $(1-\tau)$ evaporates en route; thus if goods are shipped from, say, region 1 to region 4, only a fraction τ^3 arrives.

What will long-run equilibrium look like in this model? Obviously one possibility is that manufactures production will be evenly spread among the six regions. A second possibility is the reverse: that all manufactures production will concentrate in a single "metropolis". In Figure 4, we suppose that region 1 becomes the metropolis, its role indicated by the shading of its circle.

But there are now intermediate possibilities. Consider in particular the case illustrated in Figure 5 (ignoring the dashed arc for the moment). In this case there is one metropolis in region 1, but a second one in region 4. Each of these metropolises has a "hinterland" of two rural regions: while they will sell manufactures into each other's hinterland, each will have a larger market share in the local area.

Evidently which kind of equilibrium develops depends on the parameters of the economy. Very low transport costs, etc., will lead to the case shown in Figure 4; very high costs to a dispersed manufacturing sector; intermediate parameters to an intermediate case.

This is still only a caricature of realistic economic geography, but it is already rich enough to shed some interesting new light on an old question. What are the effects of economic integration, especially when a small country integrates with a large one? Neoclassical economists have traditionally invoked the idea of gains from trade in both goods and factors, while critics,

from Graham (1923) on, have worried that the small country will be crowded out of increasing returns sectors. The discussion has been made vague and confusing by both uncertainty about how to model increasing returns, and about the extent to which external economies are national as opposed to international in scope.

The model sketched out here suggests a new way of thinking about this issue. A small country does not, in general, consist of small regions; it consists of fewer regions. When it integrates with a larger economy, the question is how these new regions fit into the emergent economic geography.

Consider Figure 5 again. Suppose that the six regions consist of two countries -- one comprising regions 1,2,5,6, the other comprising regions 3 and 4. (The broken line indicates the border). We suppose initially that political restrictions on trade and factor mobility are sufficient that each economy's regional structure evolves independently. Specifically, the large country develops a metropolis in region 1, while the small country develops one in region 4.

Now suppose that the countries engage in a "1992" that removes barriers to trade and factor mobility. What will happen? There are two possibilities. One is the Graham case, or the Canadian nightmare: the larger metropolis at region 1 attracts all manufacturing to itself, leaving the smaller country entirely rural. The other is that the case shown in Figure 5, rather than that in Figure 4, is the equilibrium. In that case the small country metropolis actually expands as a result of integration, as it gains access to its full natural hinterland.

This is hardly a complete analysis; but it suggests that many issues that are currently framed as issues of international trade should instead be viewed as issues of regional economics and economic geography.

REFERENCES

Arthur, B. (1989): "Competing technologies, increasing returns, and lock-in by historical events", Economic Journal 99, 167-183.

Burenstam Linder, S. (1961): An Essay on Trade and Transformation, New York: Basic Books.

Christaller, W. (1933): Central Places in Southern Germany, English translation by C.W. Baskin 1966, London: Prentice-Hall.

David, P. (1984): "The Marshallian dynamics of industrialization: Chicago, 1850-1890", mimeo, Stanford.

Dixit, A. and Norman, V. (1980): Theory of International Trade, Cambridge: Nisbet.

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

Graham, F. (1923): "Some aspects of protection further considered", Quarterly Journal of Economics 37, 199-227.

Helpman, E. (1981): "International trade in the presence of product differentiation, economies of scale, and monopolistic competition", Journal of International Economics 11, 305-340.

Helpman, E. and Krugman, P. (1985): Market Structure and Foreign Trade,

Cambridge: MIT Press.

Hirschmann, A. (1958): The Strategy of Economic Development, New Haven: Yale University Press.

Hoover, E.M. (1948): The Location of Economic Activity, New York: McGraw-Hill.

Isard, W. (1956): Location and Space-Economy, Cambridge:MIT Press.

Krugman, P. (1979): "Increasing returns, monopolistic competition, and international trade", Journal of International Economics 9, 469-480.

Krugman, P. (1980): "Scale economies, product differentiation, and the pattern of trade", American Economic Review 70, 950-959.

Krugman, P. (1981): "Intraindustry specialization and the gains from trade", Journal of Political Economy 89, 959-973.

Krugman, P. (1989): "History vs. expectations", NBER Working Paper # 2971.

Lösch, A. (1940): The Economics of Location (English translation 1954), New Haven: Yale University Press.

Lucas, R. (1988): "On the mechanics of economic development", Journal of Monetary Economics 22, 3-42.

Murphy, Shleifer, A. and Vishny, R. (1989a): "Industrialization and the big push", Journal of Political Economy, 97, 1003-1026.

Murphy, Shleifer, A. and Vishny, R. (1989b): "Income distribution, market size, and industrialization", Quarterly Journal of Economics, 104, 537-564.

Myrdal, G. (1957): Economic Theory and Underdeveloped Regions, London: Duckworth.

Perroux, F. (1950): "Economic space, theory and applications", Quarterly Journal of Economics, 64, 89-104.

Romer, P. (1986): "Increasing returns and long-run growth", Journal of Political Economy 94, 1002-1037.

Romer, P. (1987): "Growth based on increasing returns due to specialization", American Economic Review, 77, 56-62.

Rosenstein-Rodan, P. (1943): "Problems of industrialization of Eastern and South-eastern Europe", Economic Journal 53, 202-211.

von Thünen, J.H. (1826): The Isolated State (English edition 1966), Oxford: Pergamon.

Vernon, R. (1966): "International trade and international investment in the product cycle", Quarterly Journal of Economics

Weber,A. (1909): The Location of Industries (English edition 1929), Chicago: University of Chicago Press.

Young, A. (1928): "Increasing returns and economic progress", Economic Journal 38, 528-539.

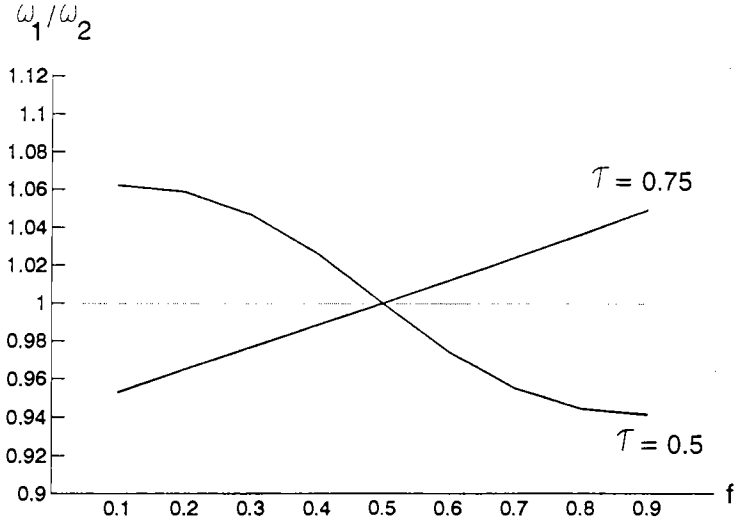
ω_1/ω_2
1.12
1.1
1.08
1.06
1.04
1.02
1
0.98
0.96
0.94
0.92
0.9
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
f
$\tau = 0.75$
$\tau = 0.5$

FIGURE 1

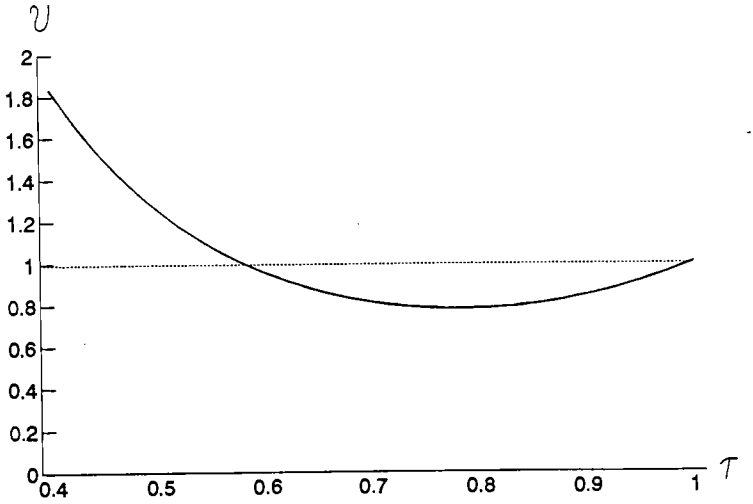

FIGURE 2

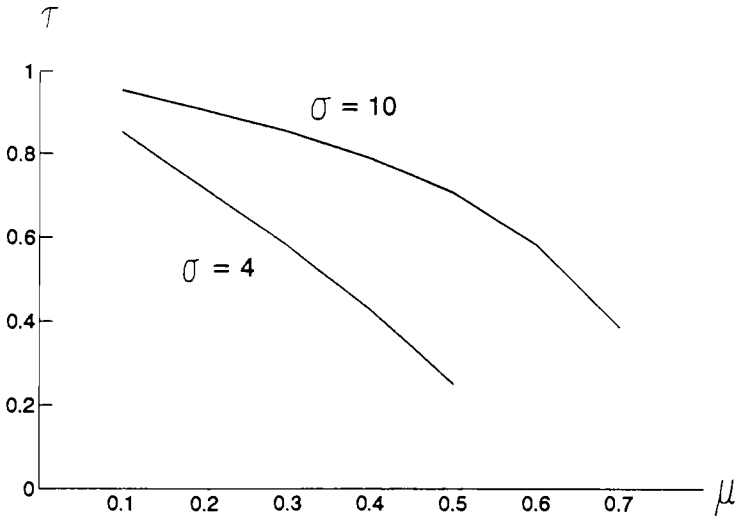

FIGURE 3

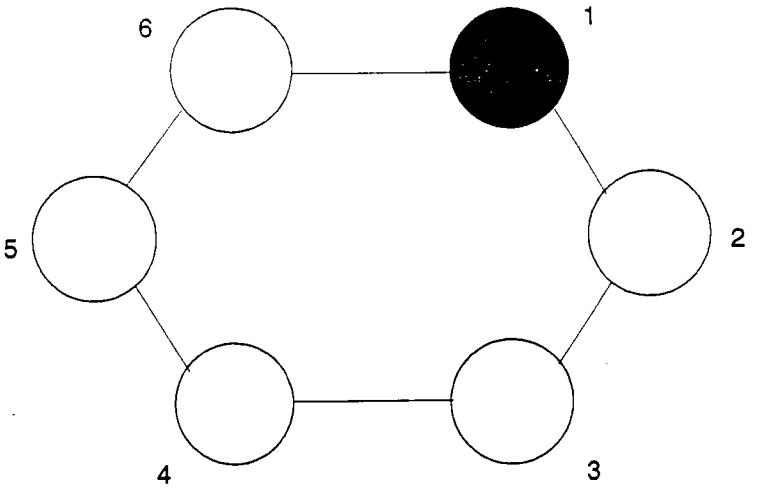

FIGURE 4

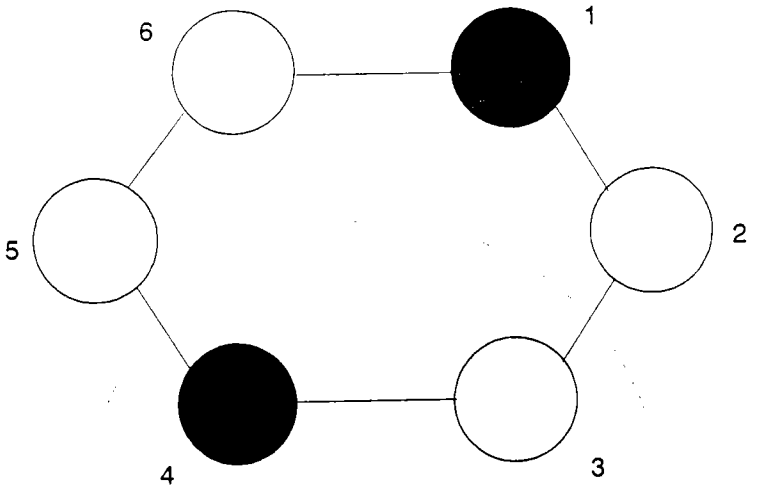

FIGURE 5