Urban Studies, Vol. 32, No. 3, 1995 563–577

Rethinking the Economics of Location and Agglomeration

Philip McCann

[Paper first received, November 1993; in final form, June 1994]

Summary. Fundamental problems exist with the classical characterisation of agglomeration economies, since such definitions do not reflect the various cost issues on which firms may wish to economise. A lack of understanding of the relationship between the notions of market hierarchies and locational behaviour leads to confusion not only in applied economic interpretation, but more fundamentally in the construction of theoretical location models. In particular, neo-classical location theory can be shown to be crucially flawed as a basis for spatial analysis. This paper therefore attempts to provide an alternative definition of the various types of agglomeration economies such that the various strands of economic theory might be used in a more rigorous manner in the discussion of spatial increasing returns.

Introduction

The advent of the ‘New Growth’ theories has led to a recent revival of interest in the nature of endogenous technological change. In particular, Paul Krugman’s (1991) book *Geography and Trade* and Brian Arthur’s (1990) paper ‘Positive feedbacks in the economy’ have both re-stimulated the discussion as to exactly what role space and location play in such questions of growth. Yet, for urban and regional economists these are not new questions, since such discussions are at the very heart of the discipline. Indeed, the role which space and distance play in determining the nature and behaviour of the economy is the central departure point which defines the urban and regional economic paradigm. Here, the spatial corollary of aspatial increasing returns to scale is economies of agglomeration, and the spatial corollary of aspatial decreasing returns to scale is diseconomies of agglomeration. However, behind this terminology lie the questions as to why, when, where and under what conditions such processes should occur in space. These questions are fundamentally questions of location.

The microeconomic methodology which urban and regional economic analysis offers in order to try to answer such questions is location theory. This is a major field in its own right. Yet, when we attempt to use these theories in order to explain the phenomenon of modern real-world agglomeration tendencies, and especially those which involve medium or large-scale firms, then we are faced with two frequently observed phenomena which are very difficult to explain using existing paradigms. The first paradoxical phenomenon we often see, is that a large proportion of firms have few or no trading

Philip McCann is in the Regional Science Department, University of Pennsylvania, 130 McNeil Building, 3718 Locust Walk, Philadelphia, PA 19104-6209, USA.

0042-0980/95/030563-15 © 1995 Urban Studies

links with other local firms in the same industry, even when there is a strong spatial clustering of a particular industrial sector. As such, the validity of the notion of the importance of localisation economies becomes somewhat questionable. Secondly, a large proportion of firms have few or no trading links with other firms or households either in the same urban area or even in the same geographical region in which they are located, even though the area comprises a clustering of various economic activities. As such, the validity of the notion of the importance of urbanisation economies becomes somewhat questionable, at least as far as the location of intermediate goods suppliers and final consumer markets is concerned. These paradoxes are further reinforced by the fact that they frequently occur simultaneously. Given that there is nothing inherently spatial about increasing returns to scale *per se*,[1] then applied research frequently finds itself at something of an impasse when faced with analysing one of the many cases where spatial clustering occurs without any significant local input–output relationships. Under these circumstances, researchers may resort to discussions as to the possible importance of localised information flows. However, such spatial clustering often takes place either in industries in which the innovation rate and the speed of technological change are not high or, alternatively, in industries which also coordinate complex activities on a global scale. All of these observations therefore still leave us with the fundamental unanswered question as to why clustering should take place.

This paper will argue that the extent to which existing location theory can provide a theoretical underpinning for discussions as to the nature of real-world agglomeration economies or diseconomies of scale is rather limited. The reason for this is that underlying spatial economic questions are questions of the nature of production hierarchies. A lack of conceptual clarity on this point has led to several definitional problems concerning, first, the construction of location theory models and, secondly, the characterisation of various types of agglomeration economies—i.e. internal returns to scale, localisation economies and urbanisation economies. The result is that as soon as any theoretical location models which allow for input substitution are used as a basis for discussing real-world spatial phenomena, then we immediately run into several problematic issues of definition and meaning. This creates a further problem in that, since input substitution behaviour is the fundamental tenet of existing microeconomic theories of the firm, these conceptual problems must necessarily be overcome in order to give the firm in space a philosophical basis which is reconcilable with that of the aspatial firm. Otherwise, spatial economics will have nothing unique to contribute to the general aspatial debate as to the nature and behaviour of the firm.

It will be shown that these conceptual problems can be overcome by simply defining the input substitution behaviour of the firm in space in exactly the same way as that of the aspatial firm—i.e. by defining capital, labour and land production factors as being mutually substitutable. The variable proportions characteristics of the inputs, and how these might be related to space, can then be discussed in the standard manner without the need for similar properties also to be ascribed to purchased inputs. Since each of the above issues will itself be associated with different kinds of real-world spatial costs, then the only way in which these various costs can be conceptually distinguished and analysed by location theory is by assuming that the physical input–output characteristics of any particular product are fixed. Given that the applied economic analysis of agglomerative behaviour necessarily involves discussions as to the types of firm in a cluster, then the conceptual distinction between these various spatial cost components is absolutely crucial if we are to be able to explain or predict such phenomena. Indeed, the fundamental problem inherent in the present definition of the various types of agglomeration economies is that they do not reflect the

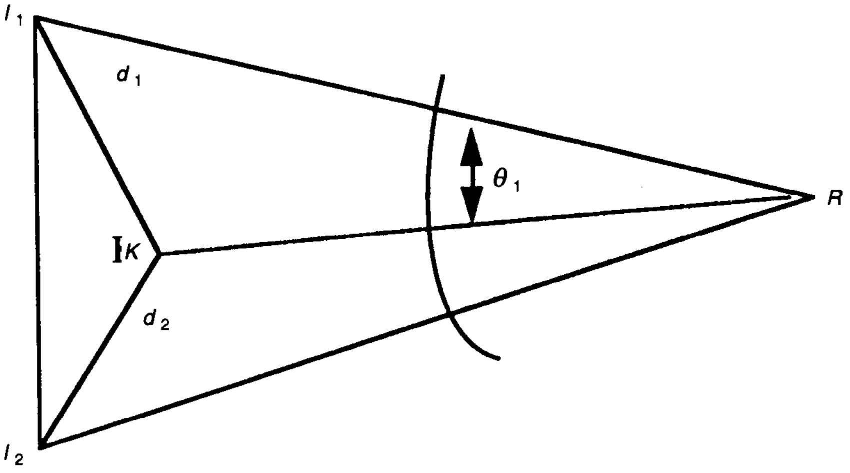

Figure 1. Location triangle.

various types of spatial costs which a firm faces. Therefore, for observed locational phenomena, the relative importance of the various cost considerations which firms will have implicitly weighed against one another and selectively economised on by their locational behaviour, may not be reflected in the resulting agglomeration characterisation. Unless the definition of both location models and economies of agglomeration follow the lines which distinguish the various kinds of spatial costs a firm simultaneously faces, then our ability to explain such phenomena will be limited. As such, not only is there very little use for the existing theoretical locational models, as they are presently defined, as an analytical basis for describing observed phenomena, but also these definitional problems themselves largely contribute to the observed paradoxes outlined above.

In order to see that this is the case, it is first necessary to re-think exactly what is meant by the construction of neoclassical location theory models.

Neoclassical Location Models: A Problem of Definition and Meaning

In 1958 Moses provided the initial fusion of location theory with production theory. Since then, many authors have developed this approach along similar lines.

The general methodology can be described as follows. It is possible to set up a Weberian locational triangle I_1,I_2R as in Figure 1, which represents the simplest two-dimensional spatial configuration within which a comparative static analysis could be achieved. I_1 and I_2 are the exogenously fixed spatial points from which a firm may purchase inputs M_1 and M_2 respectively, and R is the exogenously defined fixed market point of consumption of the product M_3 produced or to be produced by the firm in question, K. The market point could be a single firm, or an agglomeration of firms/consumers, and *fixed* in this sense implies that any locational movement of the firm K is not simultaneously associated with a locational change in the firm's customer or suppliers. Any such movement would take place in subsequent time periods, and as such, the model characterises a particular point in time, of what may be an evolutionary process.

We define the source prices of M_1 and M_2 shipped from I_1 and I_2 to K as c_1 and c_2 respectively. We assume that transport technology is unchanging—i.e. we assume that for any given material load being carried over any given distance the per ton-mile transport rate remains constant. (This does not imply that transport rates are linear with haulage distance or haulage weight.) We also assume that $M_3 = \mathrm{f}(M_1, M_2)$ where M_1 and M_2 are continuously substitutable. The two problems are:

(1) For constant relative input prices, under what conditions will a firm's optimum location be independent of the level of output?
(2) How is the optimum location of K affected by relative changes in the source prices c_1 and c_2 of the products M_1 and M_2?

The mathematical treatment of such problems within the neo-classical locational paradigm is well understood. On the first point, the general conclusion here is that where transport rates are linear or are concave with distance, then the solutions provided to these problems depend solely on the production function of the firm being homogeneous of degree one (Khalili *et al.*, 1974; Miller and Jensen, 1978). Where transport rates are concave with haulage weight, then the marginal productivity of the inputs also becomes an issue (O'Brien and Shieh, 1989; Olsen and Shieh, 1990). Meanwhile, on the second point, the solutions here will depend not only on the nature of a firm's production function, but also on the elasticity of transport rates with respect to both haulage distances and haulage volumes.

There is, however, a fundamental philosophical problem of the extent to which such an approach is able to tell us anything whatsoever about observed phenomena in real-world space. The reason is that the problem as set has little real-world economic meaning. This needs to be explained.

A firm can be defined as a production entity. A firm produces outputs by the activity of combined production factors and, in aspatial theories of the firm, the firm is defined in terms of what it does, i.e. it is defined by what it makes. This is the reference point by which individual markets are defined, since the 'supply side' of any market comprises all the firms producing, or able to produce, the good in question, and the 'demand side' are all the persons wishing to consume that particular good. Therefore, it is the nature of the good produced which allows us to discuss output prices and quantities, price elasticities and market structure. Consequently, it is the nature of the product produced which defines the relationships between individual markets, since the physical characteristics of a product will determine its complements and its substitutes. As such, a change in what a firm produces defines whether it has now moved into a different market. Similarly, where firms produce many different products, then changes in the range and distribution of outputs produced will result in the same thing. The reason for this is that the market for a particular good is simply a particular linkage within a hierarchical chain of value-adding and consumption. Heterogeneity of goods markets implies heterogeneity of production chains, irrespective of whether the goods are intermediate or final consumption goods (Williamson, 1975).

Spatial economic theory takes this approach one step further. Whereas for aspatial economic theory the one base question is what is produced and consumed at a linkage point, for spatial economic theory there are two base questions, since there is also the simultaneous additional question of where this linkage takes place. This is the dual problem which location theory attempts to answer.

In order to answer this dual problem it is necessary to begin with an analytical reference point which defines the problem. Whereas in aspatial economics this point is the nature of the good to be traded, in spatial economics the reference point can be either the nature of the product traded or the spatial point at which a linkage could occur. An example of the former approach is that of neoclassical and Weberian location theory within a two-dimensional space. In this case, the analytical reference point is the nature of the product produced, or to be produced, by the firm K. This is what defines which input source points I_1 and I_2, and which output consumption point R, can be included in the construction of the problem. It is irrelevant where these points actually are in space, in that the analytical procedure will still be able to resolve the problem of the optimum location of the firm. All that is required for the theoretical spatial problem to have an economic meaning, is for the points I_1 and I_2 to be defined as being immediately below K in the production chain, and for the point R to be immediately above K in the production chain. Assuming that I_1 and I_2 cannot produce the good M_3, then there is no direct linkage between I_1 or I_2 and R. As such, without the firm K there is a missing market for the good M_3 at R. Therefore, the very rationale for the existence of the firm K is that this would fill this missing market at R by setting up two intermediate markets at K; i.e. a market for M_1 between $I_1$and K, a market for M_2 between I_2 and K. The firm K can then set up a market for M_3 between K and R. The existence of K consequently completes this particular consumption chain in space, and the problem of finding the optimum location for K is therefore the problem of not only setting up the particular product chain, but also of maximising its efficiency by adjusting its configuration in space.

The nature of the good traded is also the fundamental reference point behind the construction of the Loschian, Hotelling or von Thünen types of location model.

An example of the latter approach is where we take the spatial location as the analytical reference point, and then discuss the effect of changes in what is produced in the firm located at that spatial point. This approach is the location theory equivalent of the trade-theory question, i.e. which good should a particular region specialise in producing. In other words, we assume factor mobility between sectors, but factor immobility between geographical regions. Then we combine information on the cost of local factor inputs and the product market price, with information on the location of markets and competitor firms in the case of each particular product. Space is the analytical reference point, and the product hierarchy is chosen so as to maximise the efficiency of the particular spatial configuration. In this particular case, although we analyse different potential product regimes, i.e. different potential markets and hierarchical production and consumption chains, the reason that we know that we are still discussing the same firm is that the firm has been defined as being situated in space, rather than in a particular product hierarchy.

In spatial economic analysis, the permanent existence of either one of these base questions, *what* or *where*, is always necessary for the construction of a model to explain *why* observed phenomena occur, and the philosophical problem with neoclassical location theory is that this fundamental requirement of an analytical reference point is not fulfilled. The model is violated by its own assumptions, such that although it makes sense mathematically, it has almost no real-world economic meaning. The problem centres around the neoclassical location theory definition of a production function.

In Moses' original paper, a change in the relative source prices of the inputs M_1 and M_2 will induce a change in the proportions of each input consumed and a simultaneous change in the location of the firm towards one input point and away from the other point. Similarly, as the firm purchases a greater quantity of input in response to an expansion in its own output demand, if the production function optimum relationships between M_1 and M_2 change for constant input source prices, then the same thing will happen. Yet, behind this paradigm is the implicit assumption that changing the relative proportions of inputs M_1 and M_2 consumed will not change the nature of the product M_3. In terms of the production of physical goods, this assumption is entirely unrealistic.

Physical products are defined in terms of their physical composition and characteristics—i.e. bulk, weight, shape and material content. It is the particular combination of such attributes which distinguishes one particular product from another, and the price of a good is the value which a consumer ascribes to a unit weight of this particular combination of characteristics and attributes.

A product may have exactly the same proportionate material content as another product, although its shape may be completely different. Therefore, these two products each embody different bundles of attributes. Similarly, the shape and visual appearance of one product may be the same as that of another product, but if they are made from different materials, or different combinations of the same material, then their material properties—i.e. durability, strength, reliability—will differ. Although products can be disguised in the short run so as to appear like other products, as long as consumers are not systematically unable to discover the durability, strength and reliability properties of the good, then the conclusion remains that a change in the physical composition of a good changes the good itself and its accompanying utility-bearing attributes. Furthermore, the fact that prices vary with overall quantities consumed does not affect this principle either, since the quantity consumed is still a multiple of a particular good, embodying a particular set of characteristics. Therefore, as far as the consumer is concerned, differences in these characteristics and properties will determine the market conditions for one good from another and, consequently, it is these same characteristics and properties of the good which a particular firm produces which will determine that firm's potential suppliers and customers. The reason is that the definition of the good defines the hierarchy of value-adding and consumption in which a firm will be at that moment. It is not possible to produce a given quantity of a particular output good from continuously substitutable purchased inputs, as distinct from capital and labour production factor inputs, without changing the output good itself.

Within the simple two-input/one-output model characterised by the Weberian triangle, a simultaneous change in both the location of the firm and the relative proportions of purchased inputs consumed by the firm means that both the location of the firm *and* the product it produces have changed. Furthermore, all sideways changes in location—i.e. movements relatively towards one input source and away from another—will be accompanied by changes in the product produced, and all product changes will be accompanied by location changes. One is never without the other. Under these conditions, how do we know that the spatial points I_1 and I_2 and R are still the immediately preceding and the immediately subsequent points in the hierarchical chain of value-adding and consumption in which the firm K will find itself, and against which we can measure spatial dimensions? The simple answer is that we do not. (See Appendix 1 for possible defences of the model.) Physical goods inherently have a qualitative nature, and to define physical goods purely in quantitative terms such as weight or volume, destroys our ability to indicate what is produced. If we cannot indicate what is produced, then we cannot discuss either markets or hierarchies. There are too many variables, and the model as set becomes meaningless as a basis for discussing why we observe that particular types of firm producing particular products exhibit particular types of locational behaviour. The reason is that physical products do not display the same variable proportions characteristics as production factors. The diminishing marginal returns philosophical justification for the assumption of continuous substitution, when discussing either the productive ability of the relationship between labour inputs and capital goods, or the utility-bearing properties of consumption goods, simply does not hold when applied to purchases. This is because productive ability and consumer utility are homogeneous abstract goods, whereas purchases are heterogeneous physical goods. Abstract goods only have a quantitative dimension inherently, and a qualitative dimension can only be given to such goods by also specifying the nature of what action is done. As far as manufacturing is concerned, this is defined by distinguishing which particular goods are produced. (It is possible to apply this same kind of reasoning to the question of returns to scale. See Appendix 3.)

Apparently, the model could therefore only have any real meaning if we were to characterise M_3 as a generic good such as cars in general, or computers in general. Yet, this level of generality would then leave us with a model in which the definition of the firm is too broad with respect to the definition of space. For example, in this particular case it would be impossible to distinguish between a firm producing automobiles from ones producing aircraft, computers, artificial limbs or hearing aids, since all of these goods are largely made from the same kinds of basic material. As such, the relative input and output transport costs per ton-mile experienced by these firms will be very similar. Therefore, if the definition of the good M_3 becomes general to the level of an individual manufacturing sector, then it becomes impossible to say anything more specific than that K represents manufacturing industry in general. As far as location is concerned, we have progressed no further than the Hecksher–Ohlin theory. (See Appendix 2).

Redefining the Costs of Space and Distance

For spatial economics, the fundamental distinction it is initially necessary to make is between the costs incurred in the overcoming of space, and the costs incurred by nature of being located at a point in space. The theoretical treatment of these two types of cost issue then immediately leads to two different paradigms, based on the assumption of two different kinds of production function.

The first kind of location costs analytical paradigm, is the distance-transactions costs paradigm, which is based on the input–output production function for purchases and outputs discussed in detail above. If we assume a linear input–output function for physical goods, then we can also discuss the question of the costs of inter-firm linkages within a production hierarchy in two-dimensional space. This is the paradigm adopted by the growth pole concept of Perroux (1950). As such, the pattern of trading networks between nodal points becomes the focus of this analysis. In this distance-transactions costs paradigm, the costs which a firm must incur are the costs generated by the shipping of a well-defined standard good between points in space. Such distance costs will comprise transport costs, telecommunications costs and the costs of inter-firm executive travel. If the product and product hierarchy are unchanging, then the latter two types of cost will be generally very low relative to the material shipment costs. This is because the necessary frequency of such contacts will be low, due to the fact that well-defined legal contracts will ensure what is to be transacted. Here, the characteristics of the good being produced are the fundamental issue governing the distance–transactions costs, since it is these which determine the volumes and characteristics of the linearly-related input and output goods to be shipped over space. The analytical base question here is what is produced.

The second kind of location cost analytical paradigm is the location-specific factor cost-efficiency paradigm, which is based on the standard neoclassical production function for factor inputs. The base question in this case is where is production. This paradigm is represented by the New Urban Economics, and the efficiency-wage theories. When applied to the question of factor costs at a point in space, the costs which a firm will incur in order to produce a particular good at any particular place, are the costs of local capital, land and labour inputs.[2]

As well as distance-transactions costs and location-specific factor efficiency costs, there is a third entirely different cost issue which a firm must face when considering the costs of location and space, and which automatically leads us to a third analytical paradigm. This other issue is the question of the nature and stability of the production and consumption hierarchy in which a firm finds itself.

In the two previous analytical paradigms the models can be well-defined because it is implicitly assumed that there is a stable production and consumption hierarchy which allows us to discuss either factor productivity[3] and/or the locations of input sources and/or market points. However, in many industrial sectors, the nature of the products or service produced are continuously changing. Under these circumstances it is not possible for a firm to know or define which spatial or aspatial hierarchy it will find itself in at any point of time in the future. If the speed of an individual product and hierarchical change is faster than the time which, concomitant with a single change in product and hierarchy, a firm would need in order to identify a better new location and to purchase the site and re-train the local workforce, then the firm will be unable to adjust optimally its locational behaviour in line with such continual changes. Examples of this phenomenon are the cases of the central capital markets, the electronics industry and the high-quality garment industry. The product produced by the capital market is the funding required for a specific project. However, the particular financing arrangements will be different for almost every single project—i.e. the product produced by the industry is always changing—and this means that different syndicates of firms and individuals will need to be constructed for each case. Similarly, in the electronics industry, the fact that product life-cycles are extremely short means that products are continuously evolving. In both of these cases, production hierarchies must therefore be continuously renegotiated and syndicates set up. Face-to-face contacts are required in order to do this, because this is the only medium as yet which can sufficiently convey the richness and complexity of such informal, tacit and formal information transactions. Therefore, if the perceived opportunity costs of not being able to maintain continuous face-to-face contact with other firms and/or customers in order to allow the maximum level of negotiation and coordination of activities, always outweigh[4] the benefits of lower factor prices at alternative locations which would allow less-intense face-to-face contact due to distance, then any consideration of alternative locations for firms in such industries, other than at locations next to the other firms performing similar activities, is completely ruled out.[5] This also explains why the quality garment industry which produces goods on a customised basis tends to be so spatially concentrated.

We can define these costs as hierarchy-coordination costs, and the approach which it is necessary to use in order to discuss the locational implications of these costs as the hierarchy-negotiation costs paradigm.

The three quite distinct supply-side cost issues described here will combine to determine the extent to which economies or diseconomies of agglomeration will occur for a firm or a group of firms. These issues are not at all sectoral, in the sense of Standard Industrial Classification sectors, but are related to the nature of the products produced, and the concomitant input–output production hierarchies. If the hierarchy-coordination costs issue is clearly the most significant for a firm, then the locational behaviour of the firm can only be analysed by adopting a probabilistic approach which broadly follows the argument outlined in Alchian's (1950) classic paper. In this case, the two-dimensional optimal location models cannot be constructed in space, because the production hierarchy is not even clearly defined aspatially. On the other hand, if it is not the case that hierarchy-coordination costs are the most important cost issue facing the firm, then we can assume a certain stability in a firm's input–output relationships over space. Under these circumstances, the locational behaviour of the firm can indeed be analysed by using a two-dimensional model which combines the distance-transaction costs and the location-specific factor efficiency costs into a single model, while still explicitly distinguishing between them. This is exactly the approach of Weber and Isard (1951), although here the conclusions can be made much richer by incorporating the 'pull' of the market into the model. This can be done within a dynamic time setting, via the use of an inventory model (See McCann, 1993).

All of the discussion so far has been concerned with the supply-side cost issues which will determine whether or not firms will tend to cluster together. However, the existing literature characterises one further potential source of agglomerative behaviour, namely the location and size of the market area. The argument here is that many firms which are selling goods to household or industrial customers will need to ensure maximum market exposure, and that the grouping of such sales outlets will maximise the total consumer interest in the products available for sale—e.g. shopping malls, restaurants and rows of car showrooms.

These demand-side locational considerations can also be treated in exactly the same way as the manner by which the various supply-side cost locational cost considerations were distinguished—i.e. by discussing the hierarchical chain of production and consumption. In such cases, the final link in the hierarchical chain of production and consumption is not well-defined: it is not certain who and where is the final consumer. Since a change in final consumer implies a change in the hierarchical production and consumption chain, then even for a stable input–output production hierarchy, the firm must decide its optimum location in an environment of continuously changing production and consumption hierarchies, which as we have seen, will imply changes in the spatial problem as set. This means that the firm must address a locational problem which includes all the potential locations of all potential customers, as well as all other locational cost issues. This is clearly impossible. The uncertainties involved in such a dynamically changing hierarchical environment mean that the calculation of a definitive optimum location is not possible, and therefore the best location, *ceteris paribus*, will be that location at which the maximum number of final links in potential consumption hierarchies are likely to be coincidental. The reasoning behind this is that, otherwise, the firm will incur the opportunity cost of lost hierarchical chains (i.e. lost sales) by not being at the point of maximum coincidence. This is sales maximisation behaviour, and the location thus chosen will be in an area of population density.[6] It is clear that this form of locational problem, and the resulting behaviour of the firm, is a similar problem to that encountered in the hierarchy-negotiation costs paradigm, although in this case the hierarchy is not well-defined even though the product may be so. We can call this problem the hierarchy-coincidence opportunity costs problem, since the aim of the firm here is to minimise the opportunity costs of a lack of hierarchy coincidence. The method of analysis here would normally be referred to as sales maximisation principle, although the important point is that the question of hierarchies is still at the root of such behaviour.

We now have four fundamentally different cost and revenue issues which are associated with the location of the firm, and which will need to be weighed against one another in order for the firm to come to some sort of location decision. The first two issues, namely the distance-transaction costs and the location-specific factor efficiency costs, will always be present, and will therefore always need to be considered by the firm when making a locational decision. Either or both of the latter two issues, namely the hierarchy-coordination costs and the hierarchy-coincidence opportunity costs, may or may not exist for the firm in question—e.g. many retail firms are also continually in the process of changing their sales produce, such that the costs of both hierarchy coordination and hierarchy coincidence will become difficult to distinguish.. Where these latter two issues do not exist for the firm, or where their importance relative to the two former issues is very limited, then the optimum location decision of the firm can be analysed primarily in a classical manner. On the other hand, if either or both of these latter two cost issues do exist, and the combined importance of these issues is great with respect to the two former issues, then the locational behaviour of the firm must be discussed primarily in terms of a probabilistic paradigm. In other words, whichever types of cost issue are clearly the most important for the firm will be the cost issues on which the firm will primarily attempt to economise and, for many firms, one cost issue which reflects a clear characteristic of the product or service produced by the firm, will tend to dominate all the others. For example, steel is a non-changing heavy product requiring non-changing heavy coal inputs, such that weight and transport costs are the crucial consideration. For firms producing customised products, the product characteristic demands hierarchy coincidence—i.e. face-to-face contacts with customers. For firms producing a wide variety of goods destined for distribution warehouses, the major need for the firm will be highly-skilled workers (or low costs of training). The resulting locational behaviour of the firm reflects an implicit or explicit attempt to economise primarily on one or more particular cost components associated with location. Therefore, where such behaviour results in several firms being spatially clustered, then in order to identify why this has occurred, it is necessary to discuss which particular cost issues each firm has been attempting to economise on by such clustering, since the various agglomeration economies which each firm will have realised will reflect one or more of the four various locational cost issues outlined above.

If either or both of the former two issues, namely the distance-transaction costs and the location-specific factor efficiency costs, are the major cost concern of the firm, then it is likely that all such firms in similar positions[7] in the same industrial sector would be located at more or less the same spatial point, assuming that they must buy from the same suppliers, and sell to the same markets. This is a very common phenomenon. This can be characterised by the Weber model, whereby the major input–output trading linkages of the firm in question are frequently with firms located at a significant distance away from the location at which the firm in question is optimally situated. If firms at the same position in a particular industrial hierarchy not only have primarily the same major input–output trading linkages but also require the same kinds of labour skills, then all of these firms will optimally locate at the same spatial point, unless this clustering process itself forces up local factor prices such that local efficiency wages become uncompetitive. Examples here are steel producers, oil refineries, Nissan in Tennessee and Toyota in Kentucky, electronics manufacturers in Scotland's Silicon Glen, and most manufacturing firms producing well-defined non-changing products. Yet, what we would observe here, is that the level of their mutual input–output trading relationships will be more or less zero, unless the calculated optimum location happened to be close to either a firm's suppliers or its customers. This accounts for the paradoxes mentioned at the beginning of the paper.

If hierarchy-coordination costs are the major cost concern of the firm, then it is likely that firms in similar production sectors, although frequently at different points in the production hierarchy, will be grouped together spatially. 'Similar' in this sense implies that the various firms intend to perform some service or activity at some point in the production hierarchy of the general kind of good being produced by the sector located there, rather than the firms necessarily being in the same SIC sector. In this case, we may well observe strong local inter-firm trading relationships. On the other hand, if hierarchy-coincidence opportunity costs are the major cost concern of the firm, then it is likely that firms in a variety of production sectors, although frequently at the same point in their respective production hierarchy, will be grouped together spatially, e.g. firms selling a variety of products, or different kinds of household service supplier firms, such as car repairs, plumbing and household electrics. This will be the case where the consumers do not have well-defined contracts with particular supplier firms. Therefore, the firms lower down the production hierarchy are themselves attempting to maximise the coincidence of their own potential hierarchies by their locational behaviour. In this case, although we will not observe strong inter-firm trading relationships amongst firms at the same relative position in the hierarchy, we may well observe strong local trading linkages with firms at different points in the hierarchy.

In either of these cases where hierarchies are not well-defined, firms will tend to group together in space, in order to ensure that the appropriate information transactions can take place. However, although these phenomena are quite different from one another, they are usually both referred to as information economies of agglomeration. Furthermore, as we have already seen, many clustering situations are wrongly characterised as localisation or urbanisation economies, when the cost reason for such clustering has little or nothing to do with the location of other local firms, but rather is due to the relationship between local factor efficiency prices and the cost considerations dependent on the location of suppliers and customers in totally different regions. The result is that authors then wrongly attempt to account for this observed spatial clustering in terms of hypothesised information economies, in situations in which this is simply not appropriate, such as the assumed comparison in some of the literature between Silicon Glen in Scotland, and Silicon Valley in California. The problem here is that the existing Marshallian definition of the various types of agglomeration economies, unfortunately does not reflect the different types of costs which a firm will be attempting to economise on by its locational behaviour. Therefore, the standard characterisation of localisation and urbanisation economies are definitions which not only may include two or more very different types of cost issue, but whose particular composition of such issues may change depending on the situation. For example, either the reduced opportunity costs of hierarchy coincidence, or the reduced hierarchy-negotiation opportunity costs, which may come about through clustering, will be defined as either economies of localisation or economies of urbanisation, simply depending on the SIC definition of the businesses concerned. By using these existing definitions, not only is it not possible to distinguish exactly which of the particular hierarchy-cost economies are being realised, but also it is not possible to indicate whether one or more of these economies are being realised. What is needed is a redefinition of the various forms of agglomeration economies, so that they coincide with the various forms of cost issues on which firms will be attempting to economise when choosing their location.

Of the above four types of costs which are associated with location, the first type, namely the distance-transaction costs, can result in clustering behaviour which is independent of the characteristics of the local area. As such, the individual firm is not achieving any form of transactions agglomeration economies of scale dependent on the existence of other firms in the same area, and such clustering is purely the incidental result of optimising behaviour. We can call this phenomenon simply industrial clustering. However, if the calculated optimal location of the firm, as calculated on a homogeneous plain, results in a location close to either a firm's suppliers or customers, then we can rightly term this agglomeration economies of proximity, in that the result has been due to the economising of distance between the firms in question.

The second type of costs which are associated with location are the location-specific factor efficiency costs, and these may or may not be associated with the existing level of industrial concentration (see note 2). An example of this difference would be the case where a region which initially has a large concentration of industry and a resulting highly-skilled local labour force, then faces a severe flight of capital which leaves behind large pools of unemployed resources. Five or ten years later, new investment may find it once again economical to re-invest in the region which at this later time no longer has an existing industrial agglomeration, but which still retains a skilled labour force due to the economic history of the region. In such a situation, where many firms now locate at the same place in order to economise on such training costs, it is appropriate to refer to this phenomenon simply as local factor-efficiency clustering. Only where it is clear that the *existing* level of agglomeration is the cause of the existing factor-efficiency prices can we rightly talk about agglomeration factor efficiencies.

The latter two forms of locational costs, namely hierarchy-coordination costs and hierarchy-coincidence opportunity costs are always dependent on the existing number of the other firms and/or households at a location. If firms cluster together in order to reduce such costs, then they will indeed be realising agglomeration economies of scale, as traditionally defined. We can therefore say that such locational behaviour allows these firms to achieve agglomeration economies of coordination and market agglomeration economies, respectively.

What we see, therefore, is that there are four different types of industrial clustering behaviour, of which there are two types which *may* involve agglomeration economies, and two types which *will* involve agglomeration economies. Importantly, there are many occasions in which the former two types of clustering behaviour will not involve agglomeration economies, in as much as such behaviour will not depend on the existing size of the urban concentration. This is what accounts for the paradoxes mentioned at the beginning of the paper.

Conclusions

This paper proposes that the way we should distinguish between, and then discuss a firm's various locational costs depends on the question of the definitional stability of the products produced and the production and consumption hierarchy in which a firm will find itself. In order to do this, it is necessary to assume that a firm comprises two quite different production functions, namely a linear input–output production function for goods, and a variable proportions production function for factor inputs. Using this as the philosophical benchmark, it is then possible to use the various strands of existing microeconomic location theory to relate individually the locational behaviour of the firm, to the nature and pattern of its inter-firm trade, the nature of its information transactions, and the importance of production factor efficiencies. This approach therefore allows us to discuss the conditions under which clustering will take place in the same manner as discussing the conditions in which clustering will not take place. This is not possible using the existing Marshallian definitions of agglomeration. Furthermore, this approach also allows the previously mentioned observed paradoxes to be resolved. To quote Stahl (1991, p. 769):

> ... it appears that the classical Weber paradigm, including its major extensions allowing for input substitution ... is by now well understood. Nonconvexities in the producers' location strategies are very likely. They tend to lead to locations at input, or output market locations, or at nodes in a network of transportation routes. Thus in retrospect it appears that far too much emphasis has been paid to the characterisation of non-existing optimal solutions to the location problem, namely intermediate locations.

I disagree completely with both of these contentions. I would suggest that the underlying implications of the principle of input substitution in location models have not at all been well understood, because the locational implications of the question of hierarchy stability has not been made sufficiently clear. This leads to misunderstandings regarding agglomeration economies such that, as least as far as manufacturing is concerned, intermediate locations are indeed the norm.

Notes

1. If a single large firm can either purchase goods in bulk economic quantities, thereby reducing its own transportation costs, or alternatively, can innovate at a faster rate than smaller firms, then there is nothing inherently spatial about these phenomena as far as agglomeration is concerned, unless the size of a single firm produces local external benefits for other firms. Such benefits will be realised as economies of urbanisation and localisation for other firms, as normally defined. As such, economies of scale are one particular potential cause of agglomeration economies, along with many others (which may be historical, topographical, legal), rather than a distinctive form of agglomeration economy in its own right.
2. The initial costs of investing at a location can

also be incorporated into this production function. These are one-off costs incurred each time a relocation occurs. Costs such as labour-training or re-training and severance pay, plus real-estate services can simply be included in the imputed stream of wage and rental costs through the standard discounting techniques. The costs of such labour training and land acquisition depend primarily on the history of the local economic environment, and as such, they may or may not have been related to the phenomenon of agglomeration. If the region had a recent history of industrial agglomeration, then an existing local skilled workforce will reduce initial training costs, and consequently increase efficiency wages, *ceteris paribus*, although crowded labour and land markets will have the reverse effect. Similarly, a lack of previous industrial investment may raise these initial training costs, although long-term wages and rents may be lower. It is these initial costs which Krugman (1991) refers to as 'set-up' costs. (For most firms, however, the sum total of these costs is negligible with respect to the total running costs of the plant, over the lifetime of a plant's location at a particular place.) Distance *per se*, plays no role here, other than defining the commuting area over which labour may be acquired.

3. In the location-specific factor-efficiency paradigm, the production factor productivity for a firm is necessarily implicitly discussed with respect to a particular output, since a production function itself depends on the quantities and prices of the good produced by the factors. Different products produced will alter the firm's production function.
4. The firm will attempt to balance the marginal benefits of the withholding of information (monopoly knowledge power of a certain product) with the marginal benefits of the sharing of information (possible inclusion in a new production syndicate and hierarchy).
5. However, such industrial sectors frequently do attempt to locate any of their activities which are based on standardised products and hierarchies to alternative lower-cost locations; e.g. standardised electronics goods manufacturing operations, household personal retail banking services, and the basic dyeing and tanning of cloth and hides. In each of these cases, the stable and clear definition of the product produced and the hierarchy in which the firm carrying out these particular activities are existing over time, allowing the firm to consider the question of the optimum location within a well-defined theoretical framework. Such activities with clearly defined products and hierarchies tend to be defined as 'low-level' activities, whereas those with continuously changing products and hierarchies tend to be defined as 'high-level' activities. The resulting observation is that 'low-level' activity firms tend to exist in lower-cost peripheral regions, whereas high-level activity firms tend to be located in central regions. This phenomenon was acknowledged in the product-cycle literature (Vernon, 1966), although the crucial point about the various stages of the product cycle, is that the production hierarchy becomes more clearly defined as we move through the cycle. The fundamental point of this paper, is that such locational behaviour is neither a question of industrial sector nor of information transmission *per se*, but rather a question of hierarchical stability.
6. This form of locational behaviour then becomes a question of intra-urban location rather than a question of inter-regional location. Within this confined context, sellers will then decide how their intra-urban location affects the combined sum of input costs, transactions costs and the opportunity costs associated with not being at a point of coincidence of the maximum number of consumer hierarchical chains. An exception to this principle is the case of mail-order firms. Instead of locating at the population point, they locate wherever they regard as the most cost-efficient site, when adding up the local factor costs plus the total costs incurred in contacting potential consumers via the postal service. Here, the postal service is used as the way of ensuring the coincidence of the maximum number of hierarchical chains of production and consumption.
7. Two firms can be defined as being as at the same position in a production and consumption hierarchy if the nature of both the input products they buy and output products they produce is similar. However, if two firms only coincide in terms of either their inputs or outputs, then we can only say that the firms will be at the same hierarchical position at whichever particular linkage point we are discussing.

References

ALCHIAN, A.A. (1950) Uncertainty, evolution and economic theory, *Journal of Political Economy*, 58, pp. 211–221.

ARTHUR, W.B. (1991) Positive feedbacks in the economy, *Scientific American*, February.

HOOVER, E.M. and GIARRATANI, F. (1985) *An*

Introduction to Regional Economics, 3rd edn. New York: Alfred A. Knopf Inc.

ISARD, W. (1951) Location theory and trade theory: distance inputs and the space economy, *Quarterly Journal of Economics*, 65, pp. 181–198.

KHALILI, A., MATHUR, V.K. and BODENHORN, D. (1974) Location and the theory of production; a generalisation, *Journal of Economic Theory* 9, pp. 467–475.

MOSES, L.N. (1959) Location and the theory of production, *Quarterly Journal of Economics*, 73, pp. 259–272.

KRUGMAN, P. (1991) *Geography and Trade*. Cambridge Mass: MIT Press.

McCANN, P. (1993) The logistics-costs location-production problem, *Journal of Regional Science*, 33, pp. 427–434.

MILLER, S.M. and JENSEN, O.W. (1978) Location and the theory of production, *Regional Science and Urban Economics*, 8, pp. 117–128.

O'BRIEN, J.P. and SHIEH, Y.-N. (1989) Transportation rates, production and location in the Weber–Moses triangle: a note, *Regional Science and Urban Economics*, 19, pp. 133–142.

OLSEN, D.O. and SHIEH, Y.-N. (1990) A note on transportation rates and the location of the firm in a two-dimension, *n*-input space, *Journal of Regional Science*, 30, pp. 427–434.

PERROUX, F. (1950) Economic space; theory and applications, *Quarterly Journal of Economics*, 64, pp. 89–104.

STAHL, K. (1991) Theories of urban business locations, in: E.S. MILLS (Ed.) *Handbook of Urban Economics*. North-Holland.

VERNON, R. (1966) International investment and international trade in the product cycle. *Quarterly Journal of Economics*, 80, pp. 190–207.

WILLIAMSON, O.E. (1975) *Markets and Hierarchies*. New York: Free Press.

Appendix 1

A first alternative way of defending the model is to say that the defined continuous, or even partial, substitutibility of M_1 for M_2, and the consequent justification for choosing I_1 and I_2 as possible input sources, is that the consumer R is completely indifferent as to the actual physical make-up of good M_3, as long as it is comprised of M_1 and M_2; consequently, R is still willing to pay a price c_3 for the good M_3, whatever its composition. Yet, if the ratio of c_2:c_1 is of the order of 10 000 (such price ratios are not uncommon in modern manufacturing industry), then in this schema M_3 would be more or less 99.99 per cent composed of M_1. However, this degree of substitutibility between M_1 and M_2 implies that, to all intents and purposes, M_1 and M_2 are perfect substitutes as far as R is concerned. Therefore, the restriction that $M_3 = \mathrm{f}\ (M_1,M_2)$, where M_1 and M_2 are continuously substitutable inputs, is purely arbitrary, given its apparent fundamental importance in defining the Weberian triangle. We might just as well say $M_3 = \mathrm{f}\ (M_1)$ and $M_3 = \mathrm{f}\ (M_2)$, under which conditions there is no rationale either for the firm K as we have defined it, or the locational problem as set, to exist. Furthermore, this general conclusion is not altered even if we allow for only partial substitutibility of purchases. The reason is that under these latter conditions, the justification for the model still depends crucially on the assumption that over the domain of substitutibility, the final consumer is indifferent as to the physical make-up of the final good M_3. In other words, the various possible mixes of M_1 and M_2 combine to produce what, by definition of their being made up from a particular combination of different products M_1 and M_2, are actually heterogeneous goods. These heterogeneous goods are then treated as being perfect substitutes for one another.

Hoover and Giarratani (1985, p. 32) attempt to justify the existing model by taking the example of a steel mill, which uses either processed iron or scrap iron as metallic inputs. In this case "... it is possible to step up the proportion of scrap at times when scrap is cheap and to design furnaces to use larger proportions of scrap at loca-tions where it is expected to be relatively cheap. In almost any manufacturing process, in fact, there is at least some leeway for responding to differences in relative costs of inputs." However, both of these inputs are basically the same good. Therefore, it is not at all clear why it is necessary to assume either non-linear marginal substitutibility, or a production function which must be comprised of both inputs. This example does not indicate a substitution principle which is generally applicable to manufacturing.

Appendix 2

The same kind of problem arises if we were to say that for a single specified output good M_3, the model parameters m_1 and m_2 could represent not single specified input products M_1 and M_2, but rather simply the total weights of any materials produced by each point I_1 and I_2 which are shipped to K. In this scenario, the firms I_1 and I_2 could be able to produce a variety of goods, and the prices c_1 and c_2 would simply reflect the average source price of the (weighted) basket of goods shipped from I_1 and I_2 to K, respectively. However, this scenario also makes the model unworkable. There are two reasons for this. First, as the relative quantity of material shipped from each input source changes, then it is impossible for us to know that the weighted average source price stays constant. Secondly, the production function of the firm *K* is defined in terms of the weights of inputs consumed with respect to the weight of output produced. Yet, if these inputs are simply the aggregated weights of a composite commodity, then the characteristics of *K*'s production function depend solely on the production flexibility of the input firms. As such, *K* does not have its own independent production function, and thus has no economic meaning as a firm.

Appendix 3

For a particular product produced by a firm, how is it possible to have increasing or decreasing returns to scale in purchases, when outputs and purchases are defined in terms of weights? Certainly, if the level of technology embodied in the production process of the firm increased, then it would be possible to achieve an increase in the returns of purchases. Examples of this phenomenon would be the case of a new catalytic method of converting and combining chemicals which allows a reduced wastage of chemical inputs, or a new method of packaging inputs which reduces the weight of material to be discarded. Yet, these are not questions of returns to scale *per se*. They are technological changes associated with time and possibly firm size, if the costs of investing in such technology are high. At a point in time, such a technological change would appear as a discontinuity in the input–output production function, whereby there would be a one-off increase in efficiency. However, for any given level of production technology, the production function for material purchases can only be a linear function of each particular output good produced. Therefore, the aggregate regional production function for purchases and output is also a linear function. This is the basis of input–output analysis.