FOUNDATIONS OF SPATIAL ECONOMICS

Arnold M. Faden

1967

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, in the Faculty of Political Science, Columbia University.

TABLE OF CONTENTS

FOUNDATIONS OF SPATIAL ECONOMICS

Introduction

The study of the spatial aspects of the economy has a long history, but only in relatively recent years has it begun to come into its own as a major field of study. Walter Isard's pair of treatises* provide the latest comprehensive summaries of the state of the art. These two works offer an instructive contras6 in point of view. The first is the latest in a traditional line of approach whose major emphasis is the development of theories giving insight into and understanding of spatial phenomena.*

* Location and Space-Economy (Cambridge, Technology Press, and New York, Wiley, 1956) and Methods of Regional Analysis (Cambridge, Technology Press, and New York, Wiley, 1960).

* The major works in this line in English, in addition to Isard's, are Alfred Weber's Theory of the Location of Industries (C.J. Friedrich, ed., Chicago, University of Chicago Press, 1928); E.M. Hoover Location Theory and the Shoe and Leather Industries (Cambridge, Harvard University Press, 1937); E.M. Hoover The Location of Economic Activity (New York, McGraw-Hill, 1948); A. Lösch The Economics of Location (W.H. Woglom, translator, New Haven, Yale University Press, 1954). Of foreign works still untranslated, the only one of comparable importance would appear to be T. Palander Beiträge zur Standortstheorie (Uppsala, Almqvist och Wiksell, 1935).

The second--as the title states--concentrates on methods which can be applied to practical problems: data-processing frameworks and optimization techniques.

The present work belongs in the traditional line, in that the primary aim is insight rather than power. Though short, it is fairly comprehensive. We have tried to push back the frontiers all along the line, and in several directions: generalization of known results, tightening up of theories to modern standards of rigor, finding new interpretations for theories, and finding new theoretical representations for spatial phenomena. Major progress has been made in some directions, very little in others.

Chapter 1, "The Theory of Commuting", is pretty much independent logically of the remaining chapters. The subject matter to be represented theoretically is the spatial movement pattern of a single individual--his itinerary, in the language of Section 1.1. A simple, rather formalistic, model for the individual's preference order among itineraries is set up, and optimized by a standard application of dynamic programming in Section 1.2. Section 1.3 introduces a little "cycling" model which brings us a step closer to the commuting models which follow, and are the main concern of Chapter 1. These are dealt with in Sections 1.4 to 1.7. The term "commuting" is used here not merely in the sense of the shuttling between home and work, but in the much wider sense of any pattern of routine movements; this usage is discussed in Section 1.4. The "commuting problem" of Section 1.5 resembles the transportation problem of linear programming: given visitation frequencies at various sites, one is to find the transport cost-minimizing pattern of trip frequencies satisfying these requirements. But the "closed" character of the commuting problem introduces essential novelties. In particular, the feasibility problem, which is rather trivial for the transportation problem, becomes quite interesting, and several striking results are obtained for it. In Section 1.6, a further specialization is made. One site, the "focal point" is singled out, and all one's trips are assumed to start or finish at this site. The optimal pattern of trip frequencies is to found, to maximize a utility function of a prescribed, fairly general, form. The main application of this model comes in the last section, where it is put to use to provide an explanation--very partial, tentative, and qualified, to be sure --for some of the mysterious empirical regularities known as "gravity" formulas.

Chapter 2 is devoted to building up a conceptual framework of basic notions around which most of spatial economics is constructed. These concepts are used freely in subsequent chapters. Section 2.1 deals abstractly with distances. It seems useful to derive these from a slightly more elementary concept, direct distance. Section 2.2 discusses the possible concrete interpretations of these numbers, and the difficulties attending them. In particular, it discusses the rather stringent conditions on transport costs which are required to establish ideal distances for pairs of sites, and ideal weights for resource-bundles. A second cluster of ideas, concerning flow-patterns over space, is discussed in Section 2.3. Still a third cluster, relating to the spatial pattern of prices, is introduced in Section 2.4. Two "efficiency" postulates are introduced, and some of their consequences spelled out. These will be assumed in all subsequent chapters. Finally, Section 2.5 introduces the concept of measure, or physical distribution over space, and some important concepts constructed jointly from distances and measures.

Section 3.1 discusses the complex nature of the location problem for the individual decision-maker, and the structure of the real estate market, and suggests some useful analytic simplifications. Section 3.2 starts by deriving the site-substitution principle, which plays a fundamental role in the rest of this work, Next, the general concept of activity is defined in a spatial context, and further sub-classified by dimension. Weberian activities are those located at single points, for which land rent is of negligible locational importance. These constitute the major subject matter of Chapter 3. The site-substitution principle is applied to the location of Weberian activities within a system of market areas. In Section 3.3 the problem of the previous section is specialized to a single market area and treated abstractly. It is then generalized in a different direction, to the optimal placement of N points; this is called the headquarter location problem. In Sections 3.4 to 3.6 the headquarter location problem is itself generalized to a class of problems in which level of output, or scale, at each headquarter point is itself a variable to be determined. This class of problems includes the central Löschian model. Section 3.4 discusses various criteria that might be used to determine the unknowns, and lists a large number of concrete interpretations--called service systems--of this generalized problem. It is noteworthy that most of these are intra-urban in character, although the Löschian system is usually interpreted as inter-urban. Section 3.5 is a formal comparative statics analysis of the general problem, under the special assumptions that service levels are uniform, placement is in a hexagonal lattice pattern, and net social benefits are maximized. The basic parameters are population density and unit transport costs. Finally, Section 3.6 investigates the Löschian model itself, addressing itself mainly to the question: what portion of the plane goes unserved by any headquarter point?

A Thünen system is a pattern of land uses in which all points equi-distant from a certain distinguished point--called the nucleus--carry the same land use. Chapter 4 is devoted to the study of these systems. Section 4.1 lists a variety of situations which are approximately Thünen systems, and discusses the conditions under which these may come into existence. Two special cases are singled out for intensive study: entrepôt models and direct-linkage models. The former is characterized by the fact that trade never occurs directly between two non-nuclear points, but only between points and the nucleus. In the latter, points can trade directly with each other in a spherically symmetric fashion, such that no cross-hauling occurs. Section 4.2 studies entrepôt systems, in which land is allocated according to the ideal real estate market of Section 3.1. It is shown that, under very weak assumptions, an ordering condition on the distance of land uses from the nucleus can be established, called the weight-falloff condition. This condition is then applied to the distribution of building heights, land speculation, and the distribution of residences by income of occupant. Comprehensive qualitative conditions of the distribution of land values can also be established. Section 4.3 deals mainly with direct-linkage models, in which available land is a power function of distance from the nucleus (a situation called homogeneous access perspective), and in which land use is determined by a single overall objective function, rather than by social equilibration. Certain far-reaching results may be derived under these conditions, e.g., the ratio of land values to transport costs, and the elasticity of demand for transportation. Finally, an attempt is made to synthesize the equilibrium approach of the last section with the maximization approach of this one. It is shown that the weight-falloff condition does solve an optimization problem, and, conversely, direct-linkage maximization emulates some of the conditions of a competitive equilibrium. This last result arises from an application of Pontryagin's Maximum Principle.

Chapter 5 is devoted to three topics which do not quite fit in with previous chapters. Section 5.1 develops a simple model to explain building height, and the distribution of real estate value between lot and improvements. Section 5.2 deals with three intra-urban problem types. The first concerns the location of a "foreign trade"-oriented land use within the context of an overall entrepôt system. The second concerns the location of a headquarter point in a Thünen system having a "city-block" metric. (In the course of this discussion, a basic convexity theorem is proved which was used without proof in Section 3.3). The third is the problem of neighborhood shape and location. We merely suggest some approaches to this very complicated problem. Section 5.3 deals with a three-population conflict situation, in which criminals prey upon victims, who are protected by police. Game theory is used to determine optimal deployments, and a two-regime solution emerges bearing a certain resemblance to the urban-rural dichotomy.

To put these results in perspective, and to chart a course for future work, we will briefly mention the major topics which have been omitted from discussion. First, dynamics. It would not be quite correct to say that time is neglected. Indeed, especially in Chapter 4, the time-spread of activities is explicitly built in, and in fact essential to some of the results, such as those concerning land speculation. But all decisions are made at the beginning, and with the passage of time we get an exfoliation rather than a development. Transportation construction and migration have been all but neglected.

Except for a few passing references, uncertainty has been assumed away. Neighborhood effects have not been discussed, except for the sketchy treatment of Sections 5.2 and 5.3. Departures from perfect competition have not been discussed (except inessentially, in Section 3.6). The rich empirical regularities of spatial economics have been discussed only insofar as a basis in theory was found for them. Thus, "gravity" models were discussed in Section 1.7, but the "central place" literature has been omitted.* (We hope to fill in many of these omissions in a larger work now in progress).

* See B.J.L. Berry and A. Pred Central Place Studies (Philadelphia, Regional Science Research Institute, Bibliographic Series, I, 1961).

Financial assistance from the Committee on Urban Economics of Resources For the Future, Inc., the National Science Foundation, the Economics Department of Columbia University, the University itself, and the Earhart Foundation is gratefully acknowledged. I have benefitted from conversations with Professor von Hohenbalken of Iowa State, my wife Mary, and Professors Becker and Vickrey of Columbia University; also from the comments of groups before whom portions of this manuscript were presented: the Labor Workshop of Columbia University, a faculty seminar at Iowa State University consisting of Professors Fox, Sengupta, Maki, and Prescott, among others. Professor Fox, my Departmental Chairman, has encouraged my endeavors in all possible ways. My wife Mary gave generously of her time in relieving me of the more onerous chores of authorship.

1. The Theory of Commuting

1.1. Itineraries

The explanation of population movements may be broken down into two steps: (1) the explanation of movements for each individual, in terms of his motives, initial location, information and opportunities; (2) the aggregation of these into gross population movements, which depends on the statistical distribution of individuals by their movement-explanatory characteristics. There is an interaction between these two levels, aggregate movements affecting the individual through several channels:
(1) Aggregate movements influence the structure of prices, wages, and rentals, and the availability of facilities, over space and time, and thereby affect individual opportunities;
(2) Aggregate movements influence the diffusion of information;
(3) Individual motives may relate directly to the spatial distribution of other people; e.g. one may have a preference for associating with certain individuals or types of people, a preference for avoiding certain others, or a preference for inhabiting one size or density of community over another.

Thus a full-blown explanation will involve a rather complex interactive system. In this chapter we shall examine one fragment of this system: the explanation of individual movements. It will become clear that this is adequately difficult in itself.

The itinerary of a person is defined to be the function giving his location at any time. It thus extends--so far as the evidence goes--from the time of a person's birth (or conception) to the time of his death, but we shall refer to the function restricted to narrower time segments also as an itinerary.

To a good first approximation, a person's itinerary can be divided into a sequence of segments with the following properties: (1) on every other segment, the location is constant--this represents a sojourn by the person at the location prescribed, for the given time-interval.

(2) on the remaining segments, which represent traveling, the itinerary is such as to make the entire function continuous (see Figurel).

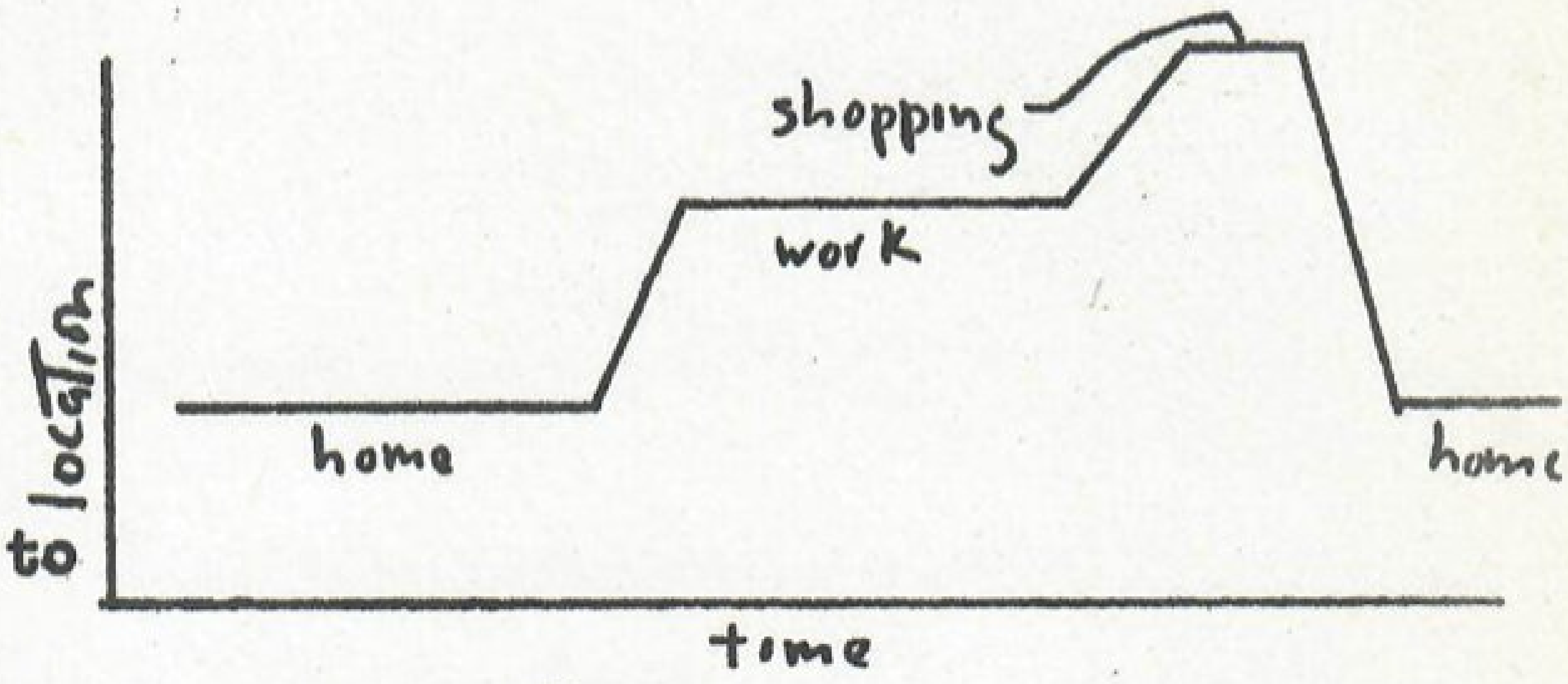

Figure 1

It is not necessary that two succeeding sojourns be at different locations. For example, one may take a walk or pleasure-drive and end up where one started.*,**

We would now like to answer such questions as: what determines length of sojourn? frequency of visits to a site? frequency of movements between sites? speed of movements between sites?

Assume a person picks the most preferred of the itineraries available to him. Preference is represented by a utility indicator which has the following form for each itinerary :

A person receives a "pay-off" during his sojourns, which accrues

*The sequence of sojourns and travels will be affected by the mode of partitioning of the landscape into "locations": the finer the partition, the more frequently will travel occur, and the shorter will sojourn times be.

**For examples of actual itineraries see P. Chombart de Lauwe, et al, Paris et l'Agglomeration Parisienne at a rate depending on calendar time t and the location L at which he is at that time, and perhaps other variables. A trip from L_1 at t_1 to L_2 at t_2 involves a cost C, depending on these four variables, and perhaps others. The utility of an itinerary is then given by

$$1) \qquad U = \Sigma_i \int_{t_i^A}^{t_i^D} V\,dt \; - \Sigma_j C_j \, ,$$

where t_i^A is the time of arrival at the i-th sojourn interval, t_i^D is the time of departure from the i-th sojourn interval, V is the "pay-off" rate; The i-summation extends over all sojourn intervals between the beginning time zero and the time horizon T; The j-summation extends over all trips taken in that period.

There are several reasonable interpretations of this formal scheme. Pay-off and cost may refer to direct satisfaction, or to income, or to some combination. (In the income interpretation, discounting is understood to be built into the C and V functions, so formula (1) need not be modified). Both C and V may assume negative values.

Pay-off rate will, in general, depend on the resources available with which one can participate in activities, at a given location and time, and the terms on which they are available. Clearly these factors vary by location; over time there are strong daily, weekly and annual cyclical components, as well as secular shifts. Furthermore, pay-off rate may depend on the length of time one has been sojourning at a given location. For example, one may have had a single purpose in visiting the site, with pay-off falling to zero after that purpose has been fulfilled: as, eating a meal, making a purchase, seeing a movie, getting an operation. More generally, a fatigue condition will be said to operate if pay-off rate eventually declines as a function of length of sojourn. An opposite pattern--which is also quite common--occurs via adaptation to local conditions, habit formation, the development of "goodwill", and "sinking roots" in a locality, all resulting in an increasing pay-off rate over time. (These two conditions are not incompatible; e.g. pay-off may be a decreasing function of present sojourn interval and an increasing function of total previous time spent at a site). Yet again, pay-off may depend on the length of absence from a site, either rising with time ("absence makes the heart grow fonder"), or falling ("out of sight, out of mind"). Pay-offs may depend on one's previous itinerary in some more complex fashion.

Clearly, the number of possible models to be explored is quite large. We will examine two rather simple ones to illustrate the possibilities. The first takes pay-off to depend only on location and calendar time. The second takes pay-off to depend only on location and sojourn length, and is of the "fatigue" variety. After that (in sections 1.4. ff.) we make further simplifications leading to many new results.

It should be noted that this entire chapter operates under the assumption of perfect information. While this could be relaxed in some of the following models, for the most part new principles would be required.

Finally, it might be mentioned that these results apply, mutatis mutandis, to other "itinerant" resources, of which the most important case is transportation equipment.

1.2. A Calendar-Time Model

Suppose one has a finite number of sites, and pay-off functions $V(L,t)$, differentiable in $\underline{t}$, defined over the sites and the time interval $[0,T]$. The transportation cost function $C(L',t',L'',t'')$ is defined on the quadruple: origin L', departure time t', destination L'', arrival time t''; $(t'<t'')$; even aside from this consideration, not all quadruples need be possible origin-destination combinations; (e.g., common carriers arrive and depart at discrete time points). One starts at some particular site at time zero, and the problem is to select the itinerary maximizing

$$2) \qquad U = \Sigma_i \int_{t_i^A}^{t_i^D} V(L,t)\,dt - \Sigma_j C(L_{j-1}, t_{j-1}^D, L_j, t_j^A).$$

The problem may be solved in two stages:

1) Out of the possible quadruples L',t',L'',t'', one selects those which give local maxima with respect to t' and t''. This can be accomplished by elementary calculus if V and C are sufficiently smooth in the time variables. For example, if for $L' = L_1$ and $L'' = L_2$ and some time neighborhood, possible crossings are constrained by $t''-t' = \bar{\theta}$, a constant, and $C(L_1, t', L_2, t'+\bar{\theta}) = \bar{C}$, a constant, The local maximum conditions are $V(L_1, t') = V(L_2, t'+\bar{\theta})$, and $\frac{\partial V(L_1,t')}{\partial t} \leqq \frac{\partial V(L_2,t'+\bar{\theta})}{\partial t}$ That is, pay-off rate at origin at departure time equals pay-off rate at destination at arrival time, and the latter is rising at least as fast as the former. If all departure and arrival times which are possible are discrete, for all pairs of locations, this step may be omitted.

2) For all problems that might arise in practice, the preceding step will result in a finite sprinkling of "reasonable crossing" quadruples. The problem then reduces to the combinatorial one of selecting that sequence of crossing quadruples yielding the best itinerary. For this, the method of dynamic programming is well adapted.*

Dynamic programming embeds the original problem in a wider class of problems of the same type. In the present case, this class is: find the optimal itinerary starting from location L at time t, for all L, and all t $\in [0,T]$. Let U(L,t) be the function giving total net pay-off (i.e., total pay-off minus total transport costs) under an optimal itinerary starting from location L at time t, and extending to the horizon T. Let $\tilde{t}$ be the earliest time satisfying the following conditions:

$t < \tilde{t}$, and there is an (L'', t''), such that $(L, \tilde{t}, L'', t'')$ is a "reasonable crossing" quadruple; if no such $\tilde{t}$ exists, we take $\tilde{t} = T$. That is to say, $\tilde{t}$ is the earliest time *after* t at which we contemplate taking a trip away from site L, and if there is no such time, $\tilde{t}$ is the time-horizon T.

Next, consider the two possible cases: (1) there is not a reasonable crossing quadruple starting from location L at time t; or (2) there is. In the first case, it must be true that

$$3) \qquad U(L,t) = \int_t^{\tilde{t}} V(L,\tau)\,d\tau + U(L,\tilde{t}).$$

*R. Bellman *Dynamic Programming* (Princeton, 1957);
R.E. Bellman and S.E. Dreyfus *Applied Dynamic Programming* (Princeton, 1962)

(3) follows from the fact that, up to time $\tilde{t}$ at least, there is no reasonable option except to remain at site L. Total net pay-off must then equal the sum of pay-off at site L up to time $\tilde{t}$, which is given by the integral, plus net pay-off after time $\tilde{t}$, which is $U(L,\tilde{t})$, by definition.

In the second case, it must be true that

$$4) \quad U(L,t) = \max \left\{ \begin{array}{l} \int_t^{\tilde{t}} V(L,\tau)\,d\tau + U(L,\tilde{t}) \\ \max\limits_i \left[U(L''_i, t''_i) - c(L,t,L''_i,t''_i) \right] \end{array} \right\}$$

This relation follows from the following considerations. In the first place, one still has the option of remaining at L. This leads to a net pay-off as in formula (3), whose right-hand side is repeated in the top line of the maximand. But one also has one or more options to cross over to a new location. We let $\underline{i}$ index these possibilities, L''_i being the i-th destination and t''_i the corresponding arrival time at that destination. The net pay-off for this option is equal to the net pay-off starting from location L''_i at time t''_i, minus the transport cost of getting to L''_i at t''_i. Finally, the option chosen for an optimal itinerary will be one having the highest net pay-off, and this maximum will be the value of U(L,t)--which is just what relation (4) states.

Thus relation (4) not only gives a functional equation for net pay-off, but also indicates the correct move to make at each choice point: viz., choose the move which maximizes the right hand side. (If several moves give the same highest value, they are all indifferently optimal).

The remaining task is to compute the optimal net pay-offs and moves. This is accomplished by the standard procedure of backward recursion.* First we note that $U(L,T) = 0$, for all L, by the definition of horizon. By substitution into (3) we determine net pay-off for the latest arrival points at every site. Substitution of these results into (4) determines net pay-offs for the latest departure points at every site. By continuing this process alternately in (3) and (4) we work our way backward in time, and ultimately determine the entire structure of net-pay-offs and optimal itineraries.

An arithmetical example. We suppose that stage one has been completed, and we are left with the following system of three sites and five reasonable crossings. (Figure 2).

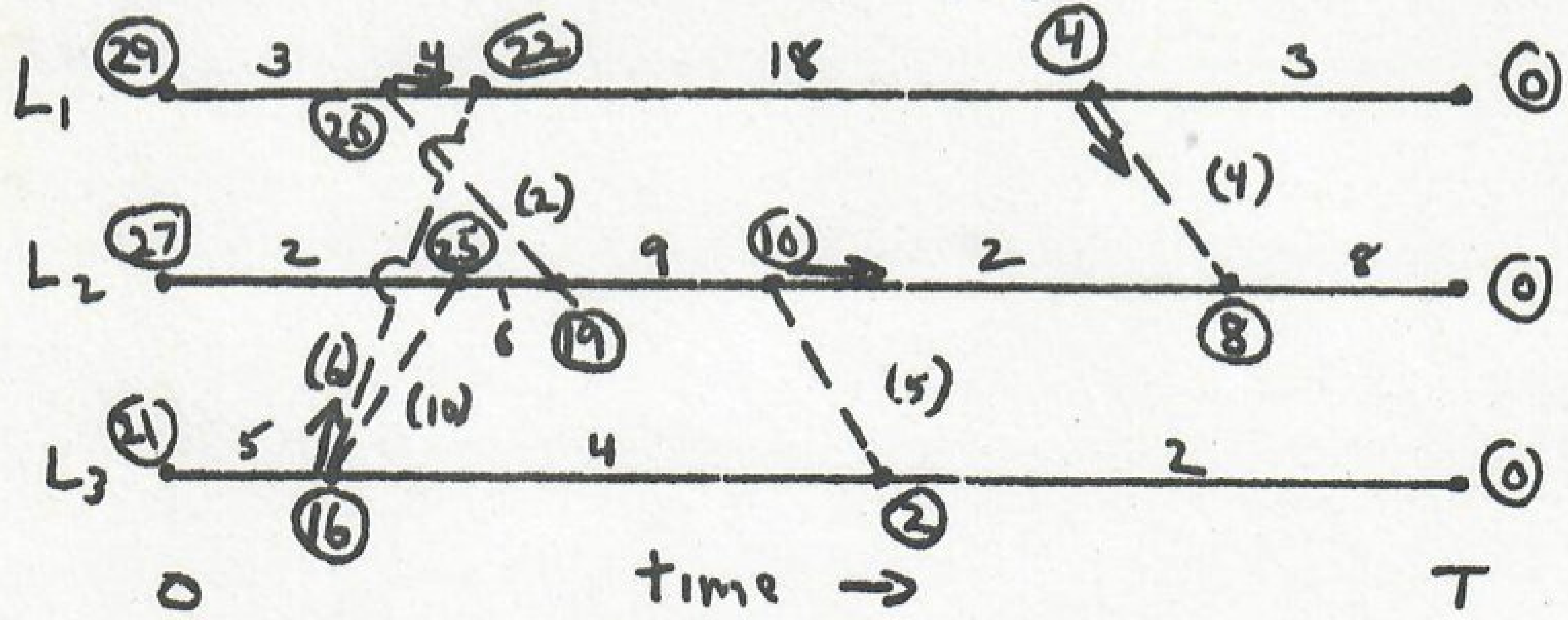

Figure 2

Reasonable origins and destinations are indicated by dots; these mark off a sequence of time-segments at each site. Each of these segments is marked with a number indicating the pay-off integral for that segment (e.g. in L_3 there are three semgents, with integrals of 5,4,2). Each dashed line represents a reasonable crossing; each is marked with a number in parentheses indicating transport cost for that trip. This is the total body of given data. (These numbers were chosen more or less at random).

*see Bellman and Dreyfus, op.cit., for numerous examples.

Now, starting at T and working leftward, we find successively U(L,t) at each of the dots (the values are circled); e.g., for the second dot from the left on L_3, we find net pay-off to be max (4+2, 22 - 6, 25 - 10), or 16; this also indicates the choice to be made at this point, viz, move to L_1 (indicated by the arrow) since this gives the highest value, 22 - 6. General conclusions may be read off the diagram: e.g. that tho best place to start out from is L_1, that someone starting from L_2 should not move at all, that everybody ends up at L_2 at time T, etc.

1.3. A"Fatigue" Model

The following simple model may be contrasted with that in 1.2. Here again we have a finite number of sites; the pay-off rate at site L_i is $V_i(t)$, where t is not calendar-time, but sojourn time: that is, the time-interval which has elapsed since one's latest arrival at site L_i. We assume that $V_i(t)$ is a continuous function which, for t large enough and for all i, decreases to such an extent that it pays to travel to another site. In fact we simply assume here that one "makes the rounds" of all sites in a regular cycle; the major question is to determine optimal sojourn length at each site. The total transport costs for a round-trip cycle is a given constant C, and the total travel time for a round-trip cycle is another given constant Q.
It is simplest in this model not to have a time-horizon; instead we take as our criterion the average pay-off rate: that is, the total pay-off over a cycle divided by the time-length of the cycle. Formally, we are to maximize

$$5)\qquad U = \frac{\Sigma_i \int_0^{\bar{t}_i} V_i(\tau)\,d\tau - C}{\Sigma_i \bar{t}_i + \theta}$$

over the $\bar{t}_i$, where $\bar{t}_i$ is the sojourn length at the i-th site.

The solution may be found by elementary calculus. The formal conditions characterizing the solution are quite simple: If we write $\bar{U}$ for the maximal attainable value of U, then

5) $V_i(\bar{t}_i) = \bar{U}$ at optimal $\bar{t}_i$, for all i.

(Also, $V_i(\bar{t}_i)$ is non-increasing at optimal $\bar{t}_i$.

This result is, in fact, intuitively evident, since if $V_i(\bar{t}_i) > \bar{U}$, we obviously raise the average pay-off rate by staying a little longer at site i; (if marginal > average, then average is rising); conversely, if $V_i(\bar{t}_i) < \bar{U}$, it pays to leave a little sooner.

It can also be shown that a rise in C increases sojourn lengths (or leaves them unchanged, as a limiting case). In effect, one spreads the extra "overhead" over a longer "production" cycle. Also, if $\bar{U} > 0$, then a rise in travel time θ also increases sojourn lengths (or leaves them unchanged).

Incidentally, this model (and all the others in this chapter, in fact) has applications outside the spatial realm, since it may be interpreted as describing an individual switching from one activity to another, rather than from one site to another. In particular, it may be useful for explaining the allocation of time among work, sleep, various leisure activities, etc.*

*on this problem, see G.S. Becker "A Theory of the Allocation of Time" Economic Journal January, 1965.

1.4. Migration and Commuting

It is often very useful to divide an activity conceptually into *current operations* and *investment*. We make an analogous distinction here between *commuting* and *migration*, respectively. In common usage, "commuting" refers to the daily shuttle between home and work.# Here it is used in the wider sense of referring to any *routine pattern of movement*. Thus any traveling cycles, any trips that one makes on a more or less regular basis, are considered parts of one's commuting routine, whether it be work trips, school trips by students, church trips by parishoners, annual vacations at a resort, "migratory" workers following the harvest, a troupe of performers touring a circuit, bus drivers running a scheduled route, or policemen patrolling their beats.

There is a certain vagueness about this characterization. This seeems to be unavoidable. On the one hand we are trying to isolate invariant aspects of an itinerary. Yet, since movement by its very nature involves a change, the invariants must be in the form of averages over time. For very short periods results based on averages are useless, and for very long periods secular changes may again render them useless. Yet for a broad middle range, description in terms of commuting models seems to be adequate, and these offer the advantage of greater analytic tractability than the more general model sketched in 1.1 above#

#This particular problem will be taken up below, in the context of Thünen systems.

#These difficulties of interpretation are by no means confined to commuting models. They enter, for example, in the case of steady-state inflow-outflow models, which are fictions even for the case of so-called "continuous process" industries; also, in a different way, in the case of stationary stochastic processes.

Formally, a <u>commuting routine</u> has the following structure: There are a finite number of sites which one visits; this set is called the <u>commuting span</u>. (For example, it may consist of one's residence, one's employment site, the homes of One's friends, the places where one shops, the places where one goes for recreation, perhaps a school, or church, or clinic, etc.).

Let x_i be the <u>frequency</u> with which one visits site i ; the commuting span is exactly those sites for which $x_i > 0$; let x_{ij} be the frequency with which one takes a trip from site i to site j; We then have the basic identities:

7) $x_i = \Sigma_j x_{ij}$, and $x_i = \Sigma_j x_{ji}$, where the summation extends over all commuting sites other than i.

(7) follows from the observation that the frequency of visits to a site equals the frequency of arrivals at that site, and also equals the frequency of departures from that site; also, every arrival must come from, and every departure go to, some other site. Specification of the commuting routine is completed by giving t_i, the average sojourn length per visit for each site i. (x and t are conveniently measured in the same time units, so that the product $x_i t_i$ is the fraction of one's total time spent sojourning at site i). We shall be concerned mainly with the x values in the remainder of this chapter.

We now come to migration. Migration is here taken to mean merely a change in commuting frequencies. One may distinguish three "degrees" of migration. What might be called weak migration occurs when some of the frequencies x_i change, but the commuting span remains the same, i.e., no commuting site is dropped and none is added. Partial migration occurs when one's commuting span changes, but not to a disjoint set; i.e., one keeps some of one's old haunts: (e.g. one changes jobs but not homes, or vice versa). Finally, full migration occurs when the commuting span turns over completely, as might occur when one makes a long-distance change of residence and employment (providing return visits do not occur, or that they are omitted from one's commuting span if they do occur). Note that a sequence of partial migrations may result in a full migration, if we compare only the first and last commuting span.

We might also distinguish out-migration--dropping a commuting site,-from in-migration--picking up a commuting site. The out-migration trip is the last trip away from a site one is dropping. The in-migration trip is the first trip to a site one is picking up. If one makes a full migration in one fell swoop, the migration trip is simultaneously in- and out-.

One cannot simply, by looking at an itinerary, tell what the commuting routine is at a certain time, or when a migration occurs. The fitting of these models to the data is, at least in part, a matter of judgment and convenience. For example, it may be useful to simplify by excluding infrequently visited sites from the commuting span.

A person is frequently taken to be located at his residence (e.g. for Census purposes). In effect, one site in the commuting span is singled out and made to represent the whole. This simplification might perhaps be justified in terms of the large fraction of one's time spent at home, or in terms of the high frequency of visitation to this site, or in terms of its resistance to migratory shifts.* In the next section we treat all sites symmetrically. In the sections after that we simplify further by singling out one special site--which may be taken as one's residence. Which of these approaches will prove more fruitful remains to be determined.

1.5. The Commuting Problem

The itinerary problem may now be re-formulated: what commuting routine will a person choose at any time? For the rest of this chapter we ignore the possibility of migration, and restrict ourselves to the simpler equilibrium problem: what commuting routine will a person choose? Given the set of all sites in the economy, one is to choose for each site L_i average sojourn length t_i and frequency of visitation x_i, and for every pair of sites L_i, L_j the frequency of trips from L_i to L_j, x_{ij}; the x-values being constrained by the relation (7).

As in section 1.2, we assume that the utility indicator for a commuting routine is the difference between on-site pay-offs and transport costs. On-site pay-offs depend on the visitation frequencies and sojourn lengths; transport costs depend on the trip frequencies. In this section we focus on the determination of trip frequencies; in the next two sections, on visitation frequencies.

*Place of employment might be a better overall locator in many cases.

Let us treat the visitation frequencies x_i as parameters, and try to determine the trip frequencies x_{ij} from them. (Any relations we obtain will be sub-relations of a larger system determining trip frequencies, visitation frequencies and sojourn lengths simultaneously). Suppose total transport costs $= \Sigma_i \Sigma_j c_{ij} x_{ij}$, c_{ij} being the cost incurred on a single trip from L_i to L_j. (For our purposes here, we need not worry about the source of these costs, whether in monetary outlay, or time delay, or discomfort, etc.). This suggests the following problem:

$$\text{Minimize} \quad \Sigma_{i=1}^{N} \Sigma_{j=1}^{N} c_{ij} x_{ij} \qquad \text{over } x_{ij} \geqq 0$$

$$8) \quad \text{subject to} \quad \left.\begin{aligned} \Sigma_{j=1}^{N} x_{ij} &= x_i \\ \Sigma_{j=1}^{N} x_{ji} &= x_i \\ x_{ii} &= 0 \end{aligned}\right\} \quad i = 1, \dots, N$$

This will be dubbed the <u>commuting problem</u>. The constraints (8) are a slight rewording of (7). The x_i are parameters which, without loss of generality, may be assumed positive, since if, say, $x_1 = 0$, then all trip frequencies to and from site L_1 must be zero, and we may simply discard that site from consideration.

We wish to investigate not only the <u>optimality</u> problem of minimizing transport costs subject to the constraints (8), but also the <u>feasibility</u> problem of determining when non-negative solutions to the constraints (8) exist at all. In fact, the latter problem has some surprising subtleties, and the results we obtain for it seem to be more interesting than the results for the former problem.

First, optimality. The commuting problem bears a strong resemblance to the familiar transportation problem:

$$\text{Minimize} \quad \sum_{i=1}^{M} \sum_{j=1}^{N} c_{ij} x_{ij} \qquad \text{over } x_{ij} \geqq 0$$

$$\text{subject to} \quad \sum_{j=1}^{N} x_{ij} = q_i \qquad i=1,\dots,M$$

$$9) \qquad \sum_{i=1}^{M} x_{ij} = r_j \qquad j=1,\dots,N$$

where q_i, capacity at source i, and r_j, requirement at sink j, are given parameters. Comparison of the transportation and the commuting problems reveals three differences. In two ways the commuting problem is just a special kind of transportation problem: (1) the number of sources equals the number of sinks; (2) sources and sinks can be paired off in such a manner that the capacity of a source equals the requirement of its corresponding sink. If this were all, the commuting problem would reduce trivially to known results. The last line of (8), however, provides the third, new, condition: (3) no shipments (i.e. trips) are allowed from a source to its corresponding sink.

(What this transformation of the commuting problem amounts to is the following. Each site is split into two fictitions pieces, first as a trip origin (source), and second as a trip destination (sink). This explains the 1 - 1 source-sink correspondence, and the equality of requirements (incoming trip frequencies) and capacities (outgoing trip frequencies) for corresponding sources and sinks. The condition, $x_{ii} = 0$ for all i, is true by definition in the commuting interpretation).

This transformation furnishes the key for solving the commuting problem. The method used is similar to the use of artificial variables in linear programming. We simply ignore the constraints $X_{ii} = 0$; we set c_{ij} equal to some number C, for $i = j$. This is now an ordinary transportation problem which may be solved by standard methods. The solution may, however, be infeasible for the original commuting problem, since some of the x_{ii} may be positive. We now let $C \to \infty$. This forces down the x_{ii}, and the solutions approach a feasible, and optimal, solution to the commuting problem, if one exists. If the sequence of solutions to the transportation problem give transport costs going to infinity along with C, this implies that no feasible solution to the commuting problem exists.

This procedure takes care of the optimality problem in more or less satisfactory fashion. But it is a rather cumbersome approach to the feasibility problem. We would like, if possible, to have some simple necessary and sufficient conditions on the visitation frequencies x_i for there to exist a non-negative solution to conditions (8). To this task we now turn.

The matrix equivalent of the feasibility conditions (8) is rendered as follows (see Figure 3). A square matrix is given, with all row and column totals prescribed. The i-th row sum equals the i-th cloumn sum, for all i. The main diagonal is all zeroes. Under what conditions can the rest of the matrix be filled in with non-negative numbers so as to give the required sums?*

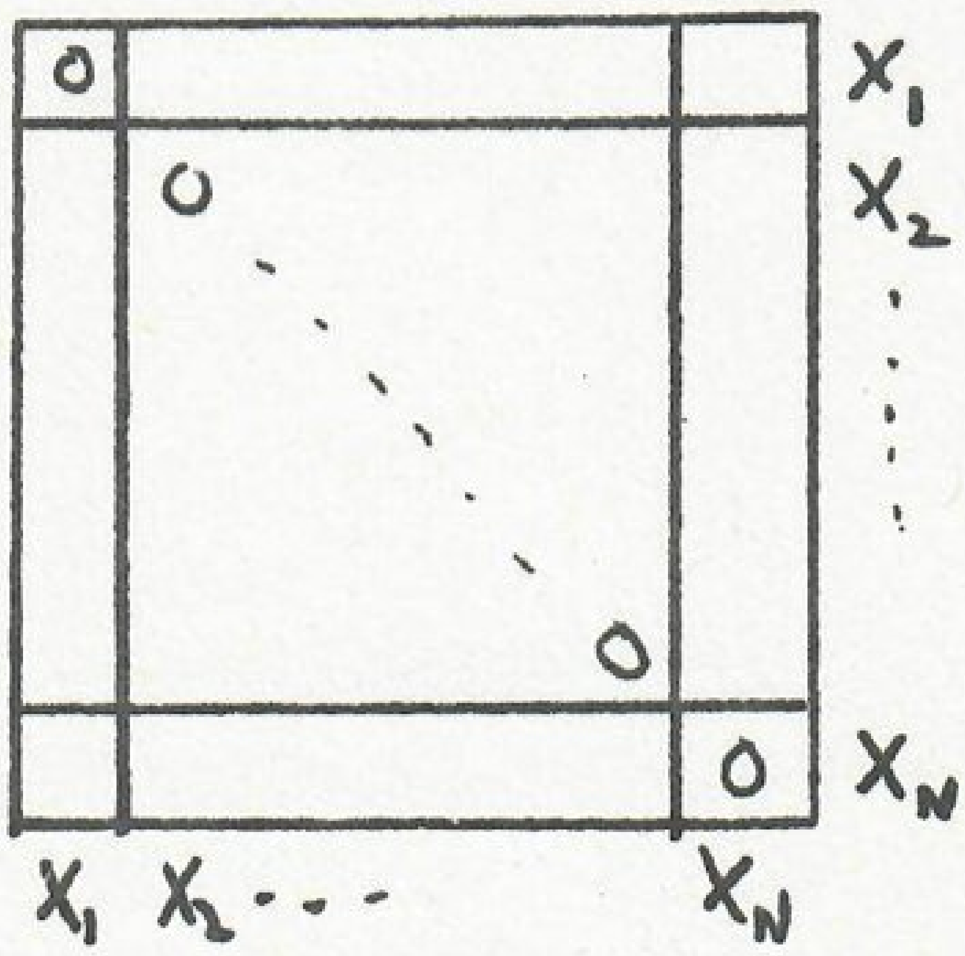

Figure 3

* The following similar problem has been investigated in the literature: The set-up is as in Figure 3, except row and column sums are not prescribed (though the i-th row sum must still equal the i-th column sum); instead, upper and lower bounds on each x_{ij} are prescribed; we are to solve the feasibility problem, and also to minimize transport costs for these new constraints. See L.R. Ford, Jr., and D.R. Fulkerson Flows in Networks (Princeton, 1962), II.3 and III.11. The methods and results seem to have little in commón with those presented here.

We say that a sequence of numbers $x_1,\ldots,x_N$ satisfies the polygon condition if and only if none of them is larger than the sum of all the others combined.*

*The name is suggested by the fact that a collection of sticks of lengths $x_1,\ldots,x_N$ can be joined to form a polygon (possibly degenerate) if and only if the x's satisfy the polygon condition.

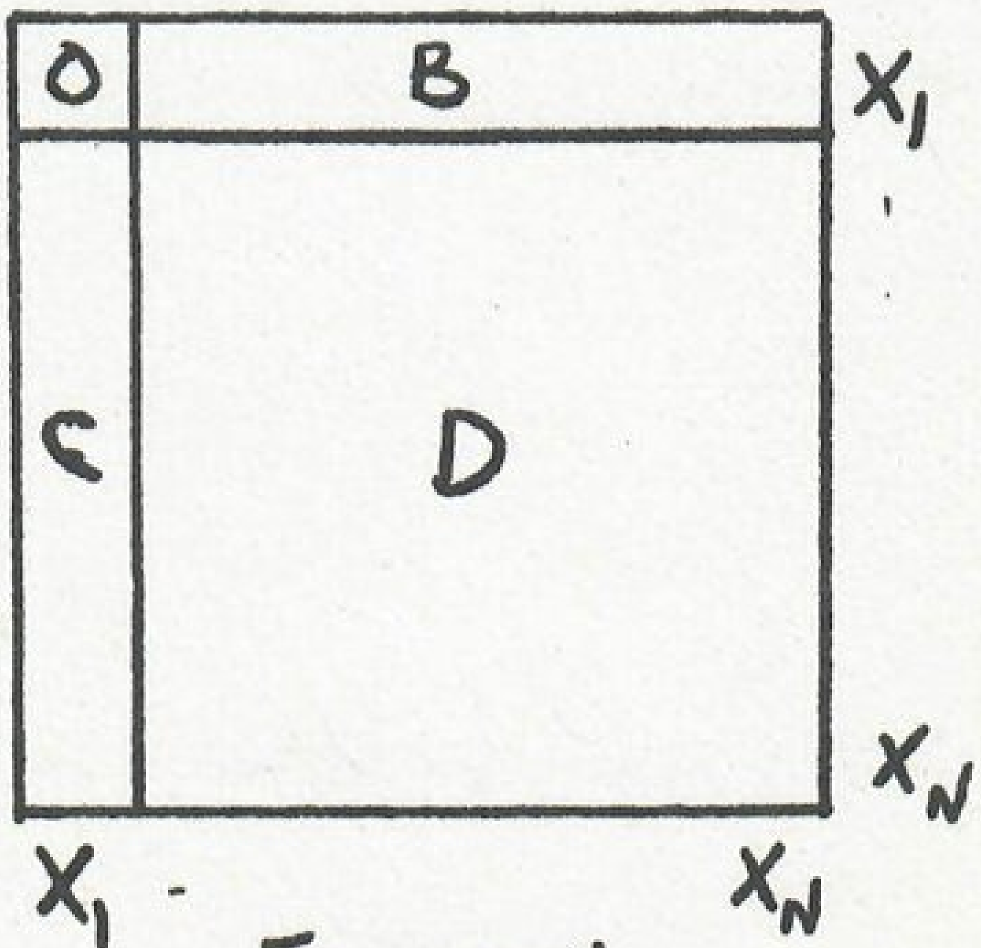

Figure 4

Theorem 1: If the commuting problem has a feasible solution, then the visitation frequencies satisfy the polygon condition.

Proof: Suppose the statement is false; then there is some sequence $x_1,\ldots,x_N$, violating the polygon condition, for which a feasible solution exists. Suppose, e.g., that x_1 is larger than the sum of all the other x's. Partition the matrix as in Figure 4, and let b,c,d be the sums of the sub-matrices B,C,D, respectively. Having regard for the row sums, we get $b > c+d$; having regard for the column sums, we get $c > b+d$; adding and canceling, we get $0 > 2d$; this is a contradiction, since the matrix D is non-negative. The same contradiction arises for any other i in place of 1. QED

This theorem is intuitively pretty obvious: for if the visitation frequency at site L_1, say, were larger than that of all other sites combined, one couldn't find other places enough to go to when leaving L_1, or enough other places to come from when entering L_1, since every trip is both a departure and an arrival.

A simple case is that of just two sites. Then one must have, for feasibility, $X_1 = X_2$ $(= X_{12} = X_{21}$, in fact$)$; which is quite obvious.

Rather more subtle is the fact that the converse of Theorem 1 is true. There may be a simple proof of this result; the only one we have found uses the theory of linear inequalities:

Theorem 2: If visitation frequencies satisfy the polygon condition, then the commuting problem has a feasible solution.

Proof: Suppose that the equations (8) do not have a non-negative solution. We now apply Farkas' lemma in the form given by Gale*.

After some simplification, we find that there exist numbers $P_1, \ldots, P_N$, and $Q_1, \ldots, Q_N$, such that

$$P_i + Q_j \geqq 0 \text{ for all } i \neq j, \quad i = 1, \ldots, N;\ j = 1, \ldots, N$$

$$\text{and } \sum_{j=1}^{N} (P_j + Q_j) X_j < 0.$$

*D. Gale The Theory of Linear Economic Models (McGraw-Hill, New York, 1960), p.44.

By this last inequality, there must be a $\bar{k}$ for which $P_{\bar{K}} + Q_{\bar{K}} < 0$, since all the x_j's are non-negative. For all $j \neq \bar{K}$, we have $p_j + Q_{\bar{K}} \geqq 0$, and also $P_{\bar{K}} + Q_j \geqq 0$; by addition, $(P_j + Q_j) + (P_{\bar{K}} + Q_{\bar{K}}) \geqq 0$; multiply this by x_j, and add over all $j \neq \bar{K}$, to get

$$\Sigma_{j \neq \bar{K}} (P_j + Q_j) x_j + (P_{\bar{K}} + Q_{\bar{K}}) \Sigma_{j \neq \bar{K}} x_j \geqq 0 .$$

To this we add the inequality $0 > \Sigma_{j=1}^{N} (P_j + Q_j) x_j$, and, after cancellation, we get

$$(P_{\bar{K}} + Q_{\bar{K}}) \Sigma_{j \neq \bar{K}} x_j > (P_{\bar{K}} + Q_{\bar{K}}) x_{\bar{K}} ;$$

$$\therefore \Sigma_{j \neq \bar{K}} x_j < x_{\bar{K}} , \text{ since } P_{\bar{K}} + Q_{\bar{K}} < 0 ;$$

but this last result shows that the polygon condition is violated.
QED

These two theorems together state that the polygon condition on the visitation frequencies is necessary and sufficient for the existence of a feasible solution to the commuting problem.

We now come to an aspect of the problem which has not yet been mentioned. The commuting routine refers to the itinerary of one person. To make sense, it must be possible to reach every site in the commuting span from every other site by using the trip frequencies in the solution. Now it can happen that a system of trip frequencies, even though satisfying the feasibility conditions (8), does not have this property. It therefore becomes of interest to find conditions under which there exist feasible solutions with this additional property as well.#

*This aspect of the commuting problem was suggested to me by Balder von Hohenbalken.

This property, called irreducibility, may be expressed in matrix form as follows. There must be no way of labelling the sites in one's commuting span that would lead to a trip frequency matrix in the form of Figure 5, where A and D are square, and O is a matrix of zeroes; for if this were the matrix, one could never get from any of the sites $L_{r+1},\dots,L_N$ to any of the sites $L_1,\dots,L_r$, since one takes a trip from a site of the former group only to another site of the same group.

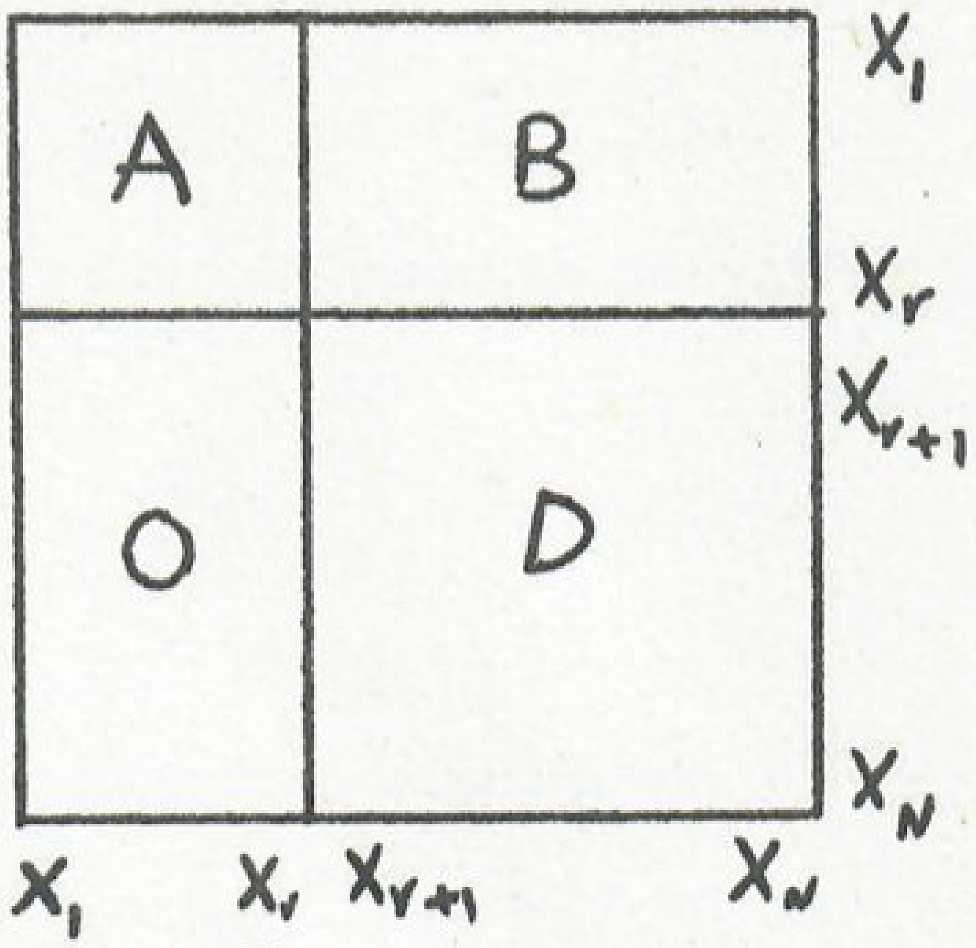

Figure 5

We first state a lemma which will be used in the main theorem on irreducibility.

Lemma: If a feasible solution to the commuting problem exists, then a symmetric feasible solution exists (i.e., one satisfying $x_{ij} = x_{ji}$, for all i,j).

Proof: If x_{ij} is a feasible solution, then $x'_{ij} = \frac{1}{2}(x_{ij} + x_{ji})$ is a feasible solution, as one verifies by direct substitution in (8); x'_{ij} is obviously symmetric. QED

This may be re-stated: if a commuting routine exists, then a bi-lateral commuting routine exists with the same visitation frequencies.

We now come to the main result:

<u>Theorem 3</u>: If a feasible solution to the commuting problem exists, then an irreducible feasible solution exists.

<u>Proof</u>: Suppose the statement is false; then there is a sequence of visitation frequencies x_1 for which there exist only reducible feasible solutions. Let N be the smallest number of sites for which such a sequence exists, and let $x_1,\ldots,x_N$ be such a sequence, labelad so that a given feasible solution has a matrix in the form of Figure 5. Pick out a symmetric feasible solution (which always exists, according to the lemma). For Figure 5 to be a symmetric matrix, the sub-matrix B must be all zeroes, too, in addition to the sub-matrix 0; Thus, the trip-frequency matrix is block-diagonal, with blocks A and D. It follows that the matrix A alone is a feasible solution to the commuting problem with visitation frequencies $x_1,\ldots,x_r$, and the matrix D alone is a feasible solution to the commuting problem with visitation frequencies $x_{r+1},\ldots,x_N$. By Theorem 1, each of these two sub-sequences must satisfy the polygon condition.

We now show the existence of an irreducible feasible solution, thus reaching a contradiction. In the sequence $x_1,\ldots,x_r$ we pick out the maximal element (or elements, if several of the x's are equally highest); suppose there are <u>m</u> such elements. Similarly, in the sequence $x_{r+1},\ldots,x_N$ we pick out the maximal elements; suppose there are <u>n</u> of them. Next, pick a small positive number ϵ, and subtract <u>$n\epsilon$</u> from each of the maximal elements of the sequence $x_1,\ldots,x_r$; also, subtract <u>$m\epsilon$</u> from the maximal elements of the sequence $x_{r+1},\ldots,x_N$; (note the reversal of the roles of m and n). It is easily verified that the new sequences resulting from these subtractions still satisfy the polygon condition, if ϵ is sufficiently small. By Theorem 2, there exist feasible solutions to the commuting problems having these new sequences as visitation frequencies; let $\tilde{A}$ be a solution matrix for the new sequence resulting from $x_1,\ldots,x_r$, and let $\tilde{D}$ be a solution matrix for the new sequence resulting from $x_{r+1},\ldots,x_N$. In Figure 5, replace the matrices A and D by $\tilde{A}$ and $\tilde{D}$, respectively.

Next, we replace certain elements of the all-zero matrices B and 0 by ϵ's: namely, in matrix B, wherever a row from a maximal element from $x_1,\ldots,x_r$ crosses a column from a maximal element of $x_{r+1},\ldots,x_N$, substitute ϵ for zero, and only in those places; similarly, in matrix 0, wherever a column from a maximal element of $x_1,\ldots,x_r$ crosses a row from a maximal element of $x_{r+1},\ldots,x_N$, substitute ϵ for zero, and only in those places. (A typical matrix resulting from these substitutions is shown in Figure 6, for the case m=2 and n=3; all the elements in B and 0, except the six indicated, remain zero).

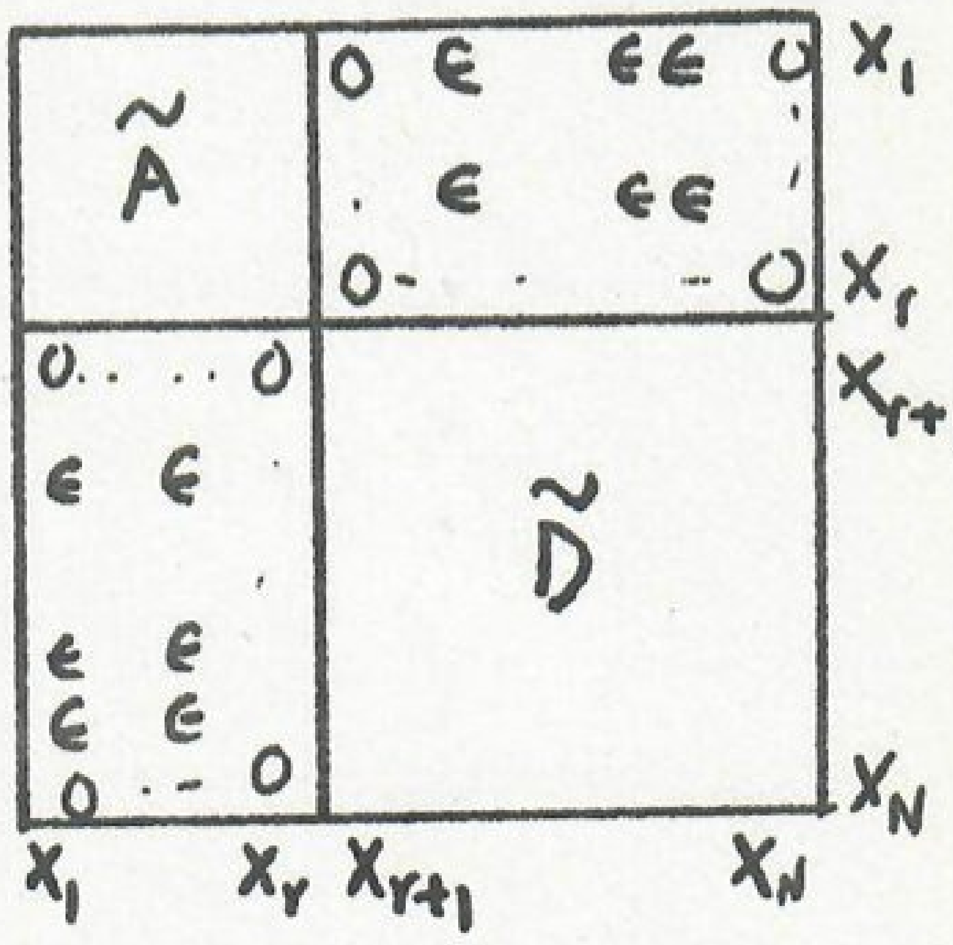

Figure 6

Both $\tilde{A}$ and $\tilde{D}$ are irreducible (or rather, can be chosen to be irreducible), since by assumption N is the smallest number of sites for which only reducible feasible solutions exist, and both $\tilde{A}$ and $\tilde{D}$ are for fewer than N sites. It is easily shown from this, and from the fact that both B and 0 are transformed to non-completely zero matrices, that the entire newly-constructed matrix is irreducible. Furthermore, it is easily verified that the row and column sums remain the same as before, the net increases in B and 0 just balancing the net decreases in A and D. We have thus reached an irreducible feasible solution and a contradiction. QED*

*It would be nice to have a short proof of this theorem.

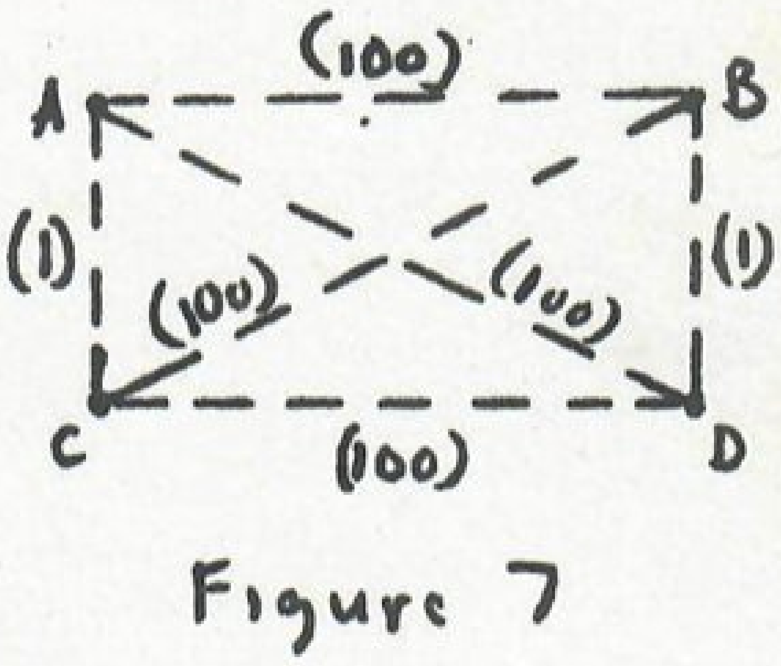

Figure 7

The significance of this last result is somewhat diminished by the fact that--even though irreducible solutions exist--the optimal solution need not be irreducible. For example, in Figure 7 we have a system of four sites with transport costs between all pairs as indicated; if all four sites have the same visitation frequencies, the obvious optimal solution to the commuting problem is to have trips only between A and C, and between B and D, which is evidently reducible. One might add further constraints to prevent this denoument, or go back to the original full itinerary formulation.

1.6. Focal-Point Models

We now restrict ourselves to the special commuting routines in which there is one site to and from which all trips are made. An example is depicted in Figure 8, a six site commuting span, in which an arc indicates that trips occur between the two sites. In the general case all possible pairs might be connected, but here we have only five connections (N - 1 in the case of N sites), and one point, called the focus, is in all of them.

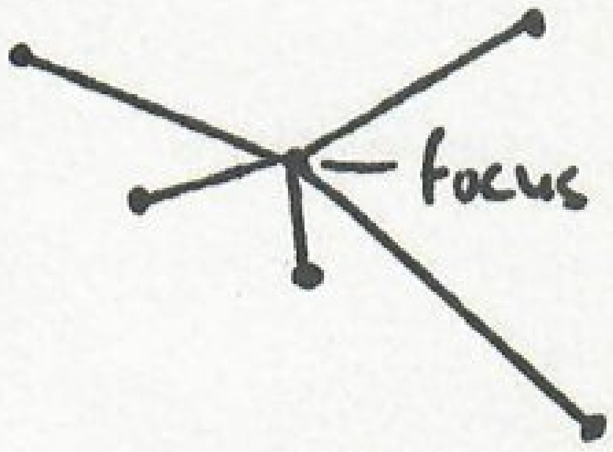

Figure 8

If the focal point, and visitation frequencies at all other sites, are specified, the commuting problem of the previous section becomes trivial, since trip frequencies are completely determined. Instead, attention will be focused on the determination of the visitation frequencies themselves. (Another problem that suggests itself is to determine the location of the focal point; this turns out to be identical with the famous Launhardt-Weber plant location problem and will be discussed in a later chapter).

The utility indicator for a pattern of visitation frequencies is assumed to be, as in previous sections, the difference between on-site pay-offs and transport costs. Formally, the problem is to

10 $$\underline{\text{Maximize}} \quad V(x_1,\ldots,x_N) - \sum_{i=1}^{N} c_i x_i \qquad \text{over } x_1,\ldots,x_N \geqslant 0$$

This will be called the <u>focal point problem</u>. Here there are <u>N+1</u> sites--the focus being site L_0--but of course only N degrees of freedom. x_i is the visitation frequency at non-focal site L_i (= trip frequency between site L_i and the focus L_0 in each direction); V is the on-site pay-off function; c_i is the round-trip transport cost between sites L_i and L_0; (as in 1.5., we need not be concerned here with the interpretation of the c_i values).

The focal point model has considerable unifying power. In the first place, there are a number of diverse spatial situations all of which have the focal point model as their abstract representation:

1) In an ordinary commuting routine, one site--most likely, one's residence--may dominate to the extent of becoming a focus;

2) As hinted at above, the focus may represent a plant-site ; in this case the x's may represent rates at which inputs arrive from various source points, and V a net profitability function (with outlays on an f.o.b. basis);or some of the x's may represent rates at which outputs flow to their various markets, or rates at which employees commute from their various residences, V and the c's being adjusted to the particular interpretation;

3) Consider a police-station assigning to each neighborhood in its precinct a certain intensity of patrolling, x_i; V here is a function representing the benefits from suppression of illegal activities, or perhaps from better traffic regulation;
4) A charity-fund headquarters assigns to each house in its environs a solicitation intensity, x_i; V here is perhaps net total funds raised, exclusive of transport-communications costs.
Numerous other examples will suggest themselves to the reader.

In the second place, the focal point problem is abstractly identical to that of the profit-maximizing firm operating in perfectly competitive factor markets, and so can avail itself of the well-developed theory of this case.*

A special case of importance is one in which the non-focal sites can be partitioned into a smaller number of clusters, such that any two sites in the same cluster are <u>perfect substitutes</u>. That is to say, $V(x_1,\ldots,x_N)$ is of the form

$$\tilde{V}\left(\Sigma_{i=1}^{N_1} x_{i1}, \Sigma_{i=1}^{N_2} x_{i2}, \ldots, \Sigma_{i=1}^{N_r} x_{ir}\right),$$

where x_{ij} is the i-th member of the j-th cluster, which contains N_j sites. Obviously, for optimality one should never travel to a given site if there is another in the same cluster which is nearer (in the sense of having a smaller c_i, not necessarily in the sense of geographical distance). This result holds even if sites in the same cluster differ in on-site costs per trip (e.g. admission charges), since these can be added to, and absorbed into, unit transport costs. This paragraph anticipates the discussion of service areas.

* P.A. Samuelson <u>Foundations of Economic Analysis</u> (Cambridge, 1947), Chapter IV.

We close this section with a simple example of spatial comparative statics. Suppose transport costs per trip fall; we may expect, in a general way, that trips will become more frequent, that new sites will enter one's commuting span, and that the trip pattern will spread out in spatial extent.

Let us write unit transport costs on the round-trip between the focus and site L_i as Θc_i, Θ being a parameter, so that we are to maximize $V(x_1,\dots,x_N) - \Theta \sum_{i=1}^{N} c_i x_i$. The optimal visitation frequencies depend on the parameter Θ, and we write them as $\bar{x}_i(\Theta)$.

Theorem: If $V(x_1,\dots,x_N)$ is homogeneous of degree K, the demand for trips, as a function of the overall transport cost level, has a constant elasticity of $\frac{1}{K-1}$, the ratio of trip frequencies to various sites being independent of the cost level; expressed formally, this states that $\bar{x}_i(\Theta) = \Theta^{\frac{1}{K-1}}\bar{x}_i(1)$.

Proof: Suppose for some Θ the statement is false. Then there are trip frequencies $\tilde{x}$, such that

$$v(\tilde{x}) - \theta c\tilde{x} > v\left(\theta^{\frac{1}{k-1}}\bar{x}(1)\right) - \theta c\left(\theta^{\frac{1}{k-1}}\bar{x}(1)\right)$$

(using vector notation) $= \theta^{\frac{k}{k-1}}\left[v(\bar{x}(1)) - c\bar{x}(1)\right]$, by the homogeneity of V, $\geqq \theta^{\frac{k}{k-1}}\left[v(x) - cx\right]$, by the optimality of $\bar{x}(1)$, for any x; in particular, for $x = \theta^{-\frac{1}{k-1}}\tilde{x}$, we get $\geqq \theta^{\frac{k}{k-1}}\left[v(\theta^{-\frac{1}{k-1}}\tilde{x}) - c(\theta^{-\frac{1}{k-1}}\tilde{x})\right] = v(\tilde{x}) - \theta c\tilde{x}$, again by the homogeneity of V; putting this chain of inequalities together, we get the contradiction

$$v(\tilde{x}) - \theta c\tilde{x} > v(\tilde{x}) - \theta c\tilde{x}$$

QED

It is interesting that we get here neither an addition to nor subtraction from one's commuting span with changing transport costs, nor, of course a change in the spatial spread. This perhaps indicates a limitation of the homogeneity assumption.

The reasonable values of the homogeneity degree are between 0 and 1, and for this range the demand is never inelastic.

1.7. A Tentative Explanation for Some "Gravity" Models

Investigators of spatial phenomena have found some remarkable empirical relations, perhaps none more so than the class of "gravity" models. These are of the form

$$x_{ij} = \alpha s_i^{\beta} s_j^{\gamma} d_{ij}^{\delta} , \tag{11}$$

where x_{ij} is a measure of the "interaction" from place L_i to place L_j (e.g. traffic, commodity flow, migration, telephone calls, mail); s_i and s_j are measures of the "sizes" of places L_i and L_j (e.g. population, total income, number of telephones, total retail sales), d_{ij} is a measure of the "distance" from place L_i to place L_j (e.g. geographic distance, monetary cost, travel time), and α, β, γ, and δ are parameters to be fitted. (β and γ are often set equal to 1). The distance exponent δ usually ranges between -1 and -3. The fits are, in general, extremely good, considering the fact that we are dealing with cross-sectional social data.*

* For a general survey, see W. Isard Methods of Regional Analysis (New York, Wiley, 1960), chapter 11.

No explanation of these regularities in terms of rational (i.e. maximizing) behavior has appeared. This is surely a challenging situation, and we offer here a small and tentative first step in this direction. The impressive feature of gravity formulae is not the "explanatory" variables themselves--"size" and "distance" variables suggest themselves naturally as predictors of "interaction"--but the power-function form in which they combine. Accordingly, we concentrate our attention on finding conditions under which a power-function will emerge from not-very-explicit assumptions as to the form of the relation.

We attack not formula (11) directly, but a somewhat weakened version of it. If "k" is substituted for "j" in (11), and the old formula is divided by the new, we obtain

12)
$$\frac{X_{ij}}{X_{ik}} = \left(\frac{S_j}{S_k}\right)^{\gamma} \left(\frac{d_{ij}}{d_{ik}}\right)^{\delta} ,$$

so that, for any two sites, L_j and L_k, the ratio of their interactions with a variable third site L_i is a power-function of the ratio of their distances from this third site.#

#Gravity relations are frequently stated in the form (12), the best-known case being "Reilly's Law of Retail Gravitation" (for which the interaction term is retail trade in shopping goods, and the exponeht δ is about -2). See W.J. Reilly "Methods for the Study of Retail Relationships" University of Texas Bulletin #2944, November, 1929 (Austin, Bureau of Business Research); O. Tuominen "Das Einflussgebiet der Stadt Turku im System der Einflussgebiete S - W Finnlands" Fennia, vol. 71, 1949.

As stated above, our main concern is to explain the power-function form of the relation (12). Suppose, then, that we drop the specific functional form in (12), and merely assume that the ratio of interactions depends on the sites L_j and L_k and on the ratio of the distances, decreasing in the latter variable:

$$\frac{X_{ij}}{X_{ik}} = f_{jk}\left(\frac{d_{ij}}{d_{ik}}\right), \quad \text{for all } L_i, \qquad (13)$$

where f_{jk} is a decreasing function. The fundamental result of this section is that, with some further, rather mild, assumptions, the non-specific form (13) actually implies the gravity model (12)!

Before we launch into details, one more general point needs discussion. The gravity models (11) and (12) are for _aggregative_ interactions, yet our analysis is an explanation of how a rational _individual_ would behave. There are two ways of justifying this procedure. (1) If all individuals have very similar preferences, the aggregative level will in some cases turn out to be an "individual writ large": that is, the functional form at the individual level is preserved under aggregation; (2) we may sometimes be justified in treating the aggregate as if it were a rational individual.* In any case, the analysis of this section can hardly be considered complete without explicitly taking account of the distribution of individuals by tastes and location, and aggregating over this distribution. This task will not be undertaken here.

Our basic model is the focal-point model (10) of section 1.6. For "interactions" we simply take the trip frequencies; for "distances" we take the unit transport cost values. In the notation of (10), relation (13) then becomes

*even sometimes if the real individuals composing it are themselves irrational. See G.S. Becker "Irrationality and Economic Theory" _Journal of Political Economy_, February 1962, pp. 1-13.

$$\text{(14)} \qquad \frac{x_j}{x_k} = f_{jk}\left(\frac{c_j}{c_k}\right).$$

Relation (14) characterizes trip frequencies as functions of the c's, which are now parameters. Site L_1 in the gravity model is the focus. The non-focal sites L_j and L_k are fixed. (Variation in the c's can come about through variation in the location of the focal point, but also by variation in fares or transport conditions between the focus, and some or all of the non-focal points).

We now make certain smoothness and simplicity assumptions about the on-site pay-off function $V(x_1,\ldots,x_N)$. None of these seems very hard to swallow.

(1) $N \geqq 3$ (i.e. there are at least three sites other than the focus); (2) V increases in each of its arguments; (3) V is thrice-differentiable;

(4) V is a strictly concave function: $V(\theta x' + (1-\theta)x'') > \theta V(x') + (1-\theta)V(x'')$ for $0<\theta<1$, in vector notation.

The basic mathematical result that we use is embodied in the following lemma.

Lemma: If $V(x_1,\ldots,x_N)$ satisfies assumptions (1) - (4), and also satisfies the condition:

$$\frac{\partial V}{\partial x_j} \Big/ \frac{\partial V}{\partial x_k} \text{ depends only on } \frac{x_j}{x_k}, \text{ for all } j,k,$$

then V is of the form $g\left(\Sigma_{j=1}^{N} a_j x_j^{\,b}\right)$ or $g\left(\Sigma_{j=1}^{N} a_j \log x_j\right)$, for some constants $a_1,\ldots,a_N$, and b, and some increasing function g, with $a_j > 0$, $0<b<1$.

Proof: (1) - (4), plus the assumption that

$$\frac{\partial V}{\partial x_j} \Big/ \frac{\partial V}{\partial x_k} \text{ depends only on } x_j \text{ and } x_k, \text{ for all } j,k,$$

implies that V is of the form $g\left(\Sigma_{j=1}^{N} h_j(x_j)\right)$, for some functions $h_1,\dots,h_N$ and some increasing function g.*

$$\frac{\partial V}{\partial x_j} \Big/ \frac{\partial V}{\partial x_k} = \frac{g' h_j'(x_j)}{g' h_k'(x_k)} = \frac{h_j'(x_j)}{h_k'(x_k)}$$ ("′" denotes differentiation); by assumption, this depends only on the ratio x_j/x_k; $\therefore \dfrac{h_j'(\lambda x_j)}{h_k'(\lambda x_k)}$ must be independent of λ, since a change in λ leaves the ratio $\dfrac{\lambda x_j}{\lambda x_k}$ unaltered; differentiate by λ, and set the result equal to zero; this yields, after some rearrangement, $$\frac{x_j h_j''(\lambda x_j)}{h_j'(\lambda x_j)} = \frac{x_k h_k''(\lambda x_k)}{h_k'(\lambda x_k)};$$ set $\lambda = 1$; since the variables are separated, and x_j and x_k can vary independently, both sides must equal a constant, $\underline{r}_j$ the resulting differential equation, $\dfrac{h_j''(x_j)}{h_j'(x_j)} = \dfrac{r}{x_j}$, on being integrated,

*Theorem 1, page 389, of S.M. Goldman and H. Uzawa "A Note on Separability in Demand Analysis" Econometrica 32:387 - 398, July, 1964.

yields $h'_j(x_j) = s_j x_j^r$, s_j being a constant of integration which, unlike r, may depend on j; a second quadrature yields

$$h_j(x_j) = \frac{s_j}{r+1} x_j^{r+1}, \quad \text{for } r \neq -1$$
$$= s_j \log x_j, \quad \text{for } r = -1;$$

(any new constants of integration appearing may be absorbed into the function g, and so are ignored); finally, set $\frac{s_j}{r+1} = a_j$ (or $s_j = a_j$, for $r = -1$), and $r+1 = b$; $a_j > 0$, from assumption (2); $0 < b < 1$, from (4). QED

We now come to the basic result.

Theorem: If a commuter maximizes $V(x_1,\ldots,x_N) - \sum_{j=1}^{N} c_j x_j$, where V satisfies conditions (1) - (4) above, and the optimal trip-frequencies, as functions of the c's, satisfy the conditions

$$\frac{x_j}{x_k} = f_{jk}\left(\frac{c_j}{c_k}\right) \qquad \text{for } j,k = 1,\cdots,N,$$

where the functions f_{jk} are decreasing, then these functionss must be of the gravity form, viz.,

$$\frac{x_j}{x_k} = \frac{p_j}{p_k}\left(\frac{c_j}{c_k}\right)^{\delta}, \qquad \text{for } j,k = 1,\cdots,N.$$

for some constants δ and $p_1,\ldots,p_N$. Furthermore, $\delta \leqq -1$.

<u>Proof</u>: The strict concavity and differentiability of V assure us that for any vector of trip frequencies x there is a cost vector c such that x is the maximizer for c, and such that the first-order conditions $\frac{\partial V}{\partial x_j} = c_j$, $j=1,\cdots,N$ hold at x.

Take two trip frequency vectors, x′ and x″, satisfying $\frac{x_j'}{x_k'} = \frac{x_j''}{x_k''}$, having associated cost vectors c′ and c″. The function f_{jk}, being decreasing, is invertible, from which it follows that $\frac{c_j'}{c_k'} = \frac{c_j''}{c_k''}$; from the first-order conditions, this is equivalent to

$$\frac{\partial V}{\partial x_j} \Big/ \frac{\partial V}{\partial x_k} \text{ at } x' = \frac{\partial V}{\partial x_j} \Big/ \frac{\partial V}{\partial x_k} \text{ at } x'',$$

so that the ratio $\frac{\partial V}{\partial x_j} \Big/ \frac{\partial V}{\partial x_k}$ depends only on the ratio x_j/x_k. We now apply the lemma, and conclude that V is of the form $g\left(\Sigma_{j=1}^{N} a_j x_j^{b}\right)$ or $g\left(\Sigma_{j=1}^{N} a_j \log x_j\right)$. Substituting these into the first-order conditions, we get

$$\frac{c_j}{c_k} = \frac{g' a_j b x_j^{b-1}}{g' a_k b x_k^{b-1}} = \frac{a_j}{a_k}\left(\frac{x_j}{x_k}\right)^{b-1},$$

or

$$\frac{c_j}{c_k} = \frac{g' a_j x_j^{-1}}{g' a_k x_k^{-1}} = \frac{a_j}{a_k}\left(\frac{x_j}{x_k}\right)^{-1}.$$

These are both in the gravity form, with $\delta = \frac{1}{b-1}$ and $p_j = a_j^{\frac{-1}{b-1}}$ in the first case, $\delta = -1$, $p_j = \frac{1}{a_j}$ in the second case; since $0<b<1$, $\delta \leqq -1$. QED

We have this derived essentially the weakened gravity formula (12) (except that we have not been able to identify separately the"size" variables s_j and the parameter γ). The assumptions made are not strong enough to go further and derive the full gravity formula (11). To do this we would have to show that the function g is linear. However, V enters only through its marginal rates of substitution, which leaves it arbitrary up to a monotone transformation (and the assumptions (2), (3), (4)), so it cannot be shown that g is linear.

The condition $\delta \leqq 1$, which drops out as a bonus, is verified in the great bulk of empirical fittings of gravity models.*

*cf. the last part of 1.6, on the elasticity of demand for trips.

Our results are narrowly limited. We have already mentioned the aggregation problem. They are explanatory only in cases where the focal-point model makes sense. For example, it is not at all clear why migration flows (as opposed to commuting) should obey a gravity law--not clear, that is, from the arguments presented in this section. Finally, the"ratio"assumptions in (13) and (14) are rather strong and somewhat arbitrary: it would be nice to derive these in turn from other plausible models.

Notwithstanding these limitations, it still has been shown that gravity models can be derived from "non-gravity-like" assumptions, and further research should be encouraged along these lines.

2. Metrics, Flow-Patterns, and Measures

This chapter will be largely definitional. We hope to set out a framework of ideas, in a more rigorous fashion than is customary in spatial economics, which will carry us through the rest of this work.

2.1 Metrics and Geodesics

In this section we give a more or less abstract presentation of leading concepts and their interrelations; in the next we will offer various interpretations of these concepts.

We suppose there are a finite number of sites in our system. We are not concerned here with the internal structure of these sites, so they may be thought of as geometrical points.
For each pair of sites (L,M) there is defined a direct distance $\delta(L,M)$. The direct distance of a site from itself, $\delta(L,L)$, is always zero. The direct distance between different sites is always greater than zero, and may be infinite for some pairs. We make no further stipulations concerning direct distances, and, in particular, we do not assume symmetry: $\delta(L,M)$ need not equal $\delta(M,L)$.

A route is simply a sequence of sites. The direct distance for the route $(L_1,\ldots,L_N)$, written $\delta(L_1,\ldots,L_N)$, is assumed to have the property

D) $$0 \leq \delta(L_1,\ldots,L_r,\ldots,L_N) \leq \delta(L_1,\ldots,L_r) + \delta(L_r,\ldots,L_N);$$

Also, $\delta(L_1,\ldots,L_N) > 0$ if not all the L's are identical.
The distance from L to M is the minimal direct distance over all possible routes beginning with L and ending with M; this is written d(L,M); it is easy to show that such a minimum always exists.

Theorem 1: d satisfies the triangle inequality:
$d(K,L) + d(L,M) \geq d(K,M)$.

Proof: Let $(K, S_1,\ldots,S_r, L)$ be a minimal direct distance route from K to L, and let $(L, S'_1, \dots, S'_{r'}, M)$ be a minimal direct distance route from Lto M. $d(K,L) + d(L,M) = \delta(K, S_1, \dots, S_r, L) + \delta(L, S'_1, \dots, S'_{r'}, M) \geqq \delta(K, S_1, \dots, S_r, L, S'_1, \dots, S'_{r'}, M) \geqq d(K,M)$; the last inequality follows from the possibility that the route shown, going through L, may not be the minimal direct distance route from K to M. QED

A route $L_1, \dots, L_N$ is a <u>geodesic</u> if and only if
$d(L_1,L_N) = d(L_1, L_2) + d(L_2,L_3) + \dots + d(L_{N-1},L_N)$.
The triangle inequality implies that the right-hand side of this expression is never less than the left-hand side for any route, so that geodesics are those special routes for which this weak inequality is pulled tight into an equality. In fact, geometrically, geodesics are simply points along a string which is pulled taut.

(It need not happen that a given geodesic remains a geodesic if the order of its sites is reversed).

<u>Theorem 2</u>: Any sub-route of a geodesic is a geodesic. (That is, if some of the sites of a geodesic are deleted, without disturbing the order of the remaining sites, then these remaining sites are also a geodesic).

<u>Proof</u>: Let $L_1, \dots, L_N$ be a geodesic, and $L_{i_1}, \dots, L_{i_r}$ a sub-route of $L_1, \dots, L_N$; $(1 \leqq i_1 < i_2 < \dots < i_r \leqq N)$; by several applications of the triangle inequality, we find
$d(L_1, L_N) = d(L_1, L_2) + d(L_2, L_3) + \dots + d(L_{N-1}, L_N) \geqq$
$d(L_1, L_{i_1}) + d(L_{i_1}, L_{i_2}) + \dots + d(L_{i_{r-1}}, L_{i_r}) + d(L_{i_r}, L_N)$
$\geqq d(L_1, L_{i_1}) + d(L_{i_1}, L_{i_r}) + d(L_{i_1}, L_N) \geqq d(L_1, L_N)$;
comparing first and last terms in this chain of inequalities,

we find that they all must be equalities;

$\therefore d(L_{i_1}, L_{i_2}) + \ldots + d(L_{i_{r-1}}, L_{i_r}) = d(L_{i_1}, L_{i_r})$;

i.e. $L_{i_1} \cdots L_{i_r}$ is a geodesic, QED

(These developments may be extended to the case of an infinite number of sites. For example, an ordered infinite set of points is defined to be a geodesic if and only if every finite sub-order of it is a geodesic. However, we have no need for this extension in this work).

The concept of "geodèsic" may now be used to define several important "betweenness" notions. Site L is between K and M iff (if and only if) the route (K, L, M) is a geodesic.

Let R' and R" be two regions (i.e., collections of sites); L is a gateway between R' and R" (in that order) iff L is between X and Y for all sites X in R', and Y in R"; L is an entrepôt for region R iff it is between all pairs of different sites, both of which are in R.

A cyclic route is one beginning and ending at the same site (and containing at least two sites). The following result is immediate.

Theorem 3: No cyclic path is a geodesic.

2.2. Metrical Interpretations

The "friction of space" which forces one to devote time, effort and resources for transportation obviously plays a fundamental role in spatial economics. The act of transportation is a rather complex affair, involving outlays, time-delay, traveling conditions, qualitative changes in the cargo, accompanying motion of vehicle and crew, with a variety of options concerning transport mode, route, speed, handling, etc. For our purposes here we simplify, and represent an act of transportation by just four variables: (1) the resource-bundle being transported (which may include people, freight, mail, water, sewage, electricity, etc.); (2) the site of origin; (3) the site of destination; and (4) the cost. There is assumed to be a function $c(b, L, M)$, determining the fourth variable from the other three. c ranges over the non-negative real numbers; L and M over the sites of the system; b over non-negative vectors giving the quantities of the resources making up the bundle being transported.

(We might instead have considered a somewhat less drastic simplification, and included also the times of departure and arrival, as in section 1.2 above. The "sites" of our system, on which our metric is to be constructed, would then be points of space-time, rather than mere locations. This procedure would have the added attraction of enabling us to treat transportation and storage simultaneously, storage being merely the special case of transportation in which the site of origin is spatially identical with site of destination. However, we stick with the simpler construction, which is adequate for our purposes).

Let us now fix on one particular $\overline{b}$ which is not the null-bundle (i.e. $\overline{b}$ contains at least some resources). This makes cost a function merely of the sites of origin and destination. The same is true of the distance function $d(L,M)$ of the previous section, and this suggests that one possible interpretation of $d(L,M)$ is as the cost function $c(\overline{b}, L, M)$ for a fixed bundle $\overline{b}$. To justify this interpretation one must show that such functions may be assumed to possess the properties ascribed to $d(L,M)$.

We started in 2.1 with the concept of "direct distance";

The direct distance $\delta(L,M)$ is now interpreted to be the cost of transporting $\bar{b}$ from L to M when one insists on using the direct route from L to M, where a direct route is one not passing through any other of our designated sites (for example, a single stretch of highway, or a non-stop plane flight). The few properties which were specified for direct distances appear quite plausible in this interpretation. In particular, if there is no direct route from L to M, so that it is impossible to go from one to the other with out passing through a third designated site, we may take the cost to be infinity, which is a permitted value for direct distance.

The actual route followed is a sequence of these direct hops, which can be represented by the sequence of sites at which the buidle successively appears, as in 2.1. The direct distance for a route is interpreted to be the cost of transporting bundle $\bar{b}$ along the route, using always the direct route between successive pairs of sites.

The second inequality of (1) states that the direct distance for a route is never greater than the sum of the direct distances for the "legs" into which it can be divided. While this inequality is not universally valid in our cost interpretation, it does reflect a pervasive feature of transportation, viz., the importance of terminal costs (e.g. waiting for connections, loading and unloading, "getting up steam", billing, information costs). Thus, when one goes straight through, using the "express" instead of the "local", one avoids terminal costs at the intermediate stopping point. As a rule, fare structures on common carriers reflect these cost differentials to the carriers.#

#Nonetheless, it is often useful to strengthen the second inequality of (1) to an equality, as will be done below.

Finally, if we assume that the route followed in going from site L to M is the least-cost route, and that this is c in the cost function $c(\bar{b}, L, M)$, we have the exact analog to the relation between "distance" and "direct distance", since $d(L,M)$ is the minimal direct distance over all possible routes from L to M. The interpretation of $d(L,M)$ as $c(\bar{b}, L, M)$, therefore, seems reasonable in many cases.

We now discuss the concept "transportation cost" in its turn. As stated above, transportation has a number of qualitatively distinct aspects; these have to be compressed into a single real number, the cost. In the case of freight transportation one can often find reasonable monetary equivalents for some of these aspects; e.g. accident risk is approximated by insurance rates; quality changes may be assessed by price comparisons; time delays are reflected in foregone interest and/or the cost of extra inventory holding. The case of passenger transportation is more difficult. We content ourselves with the observation that the trade-off rates among money, time, comfort and safety vary considerably from person to person (so that the resulting cost-metric itself varies from person to person). The problem of assessing a composite cost figure for information transmission is perhaps even more difficult.

In general, a good assessment can be made only in the context of the wider system of which the acts of transportation are a part. For example, in the calendar-time model of section 1.2, if cost (as defined there) depends only on time-delay, origin and destination, and one acts optimally, it can be shown that the marginal opportunity loss from time-delay on a trip equals the pay-off rate at destination at time of arrival (and also equals pay-off rate at origin at time of departure).

Despite these difficulties, it still seems useful to make the simplification and work with a numerical-valued (rather than vector-valued) cost function.

The metric, so far, has been defined only for the arbitrary non-null resource-bundle $\overline{b}$. If nothing else is assumed, one might get a different metric for every different resource-bundle. Suppose, however, that the cost function is factorable. c(b, L, M) is said to be factorable iff there are functions d(L,M) and w(b) which are real-valued, non-negative, for which d(L,M) is positive for L$\neq$M, for which w(b) is positive for a non-null b, and for which c(b, L, M) = w(b)d(L,M), for all b, L, M. If so, it is easy to see that substitution of one non-null bundle for another, and use of the procedure outlined above for $\overline{b}$, results in metrics that differ from each other only in a multiplicative constant, so that "essentially" the same metric is defined for all non-null bundles; this metric is proportional to the factor-function d(L,M).

These facts suggest the following procedure when we are lucky enough to have a factorable cost function. We take an arbitrary non-null resource-bundle; say, one ton of coal. The metric is defined by the cost of transporting one ton of coal. We may always choose the w(b) function so that it equals one for a ton of coal. With this, the metric is identical with the companion function d(L,M). This will be called the ideal distance from L to M. For any bundle, w(b) will be called its ideal weight. Thus, transportation cost equals ideal weight of bundle times ideal distance from origin to destination.#

*The notions of "ideal distance" and "ideal weight" date back to Alfred Weber (1909). It is curious that, while the "ideal weight" concept has been widely accepted, "ideal distances" have been given short shrift by location theorists. Weber's exposition of the latter concept may have been faulty, but it should be clear from the treatment above that the two concepts are entirely co-ordinate, and in fact are only defined jointly. For criticisms see E.M. Hoover _Location Theory and the Shoe and Leather Industries_ (Cambridge, Mass., 1937), page 40 note 10, and W. Isard _Location and Space-Economy_ (New York, 1956), page 109.

How reasonable is the assumption that the transportation cost function is factorable? There is, no doubt, a general tendency for the metrics defined by two resource-bundles to resemble each other: if L is relatively close to M for bundle $\bar{b}$, it will usually be relatively close to M for bundle $\bar{\bar{b}}$; However, there are several reasons why we may expect substantial deviations from the factorability condition.

1) Different resource-bundles may have different "comparative advantages" in being transported by different media, and these in turn may be irregularly distributed among pairs of sites. For example, suppose coal can be transported only by rail and water only by pipeline. If the pair of sites L', M' have good rail and poor pipeline connections, and conversely for the pair L'', M'', then the factorability condition breaks down.

2) A second source of deviations from factorabilty is illustrated in Figure 1, and arises from the complex character of the act of transportation which was discussed above. Suppose there is a choice of transportation modes, some fast and expensive, others

cheap but slow. For the pair of sites L',M', the curve a'b'c'd' gives the outlay-time-delay combinations which are available for the transportation of both bundles $\bar{b}$ and $\bar{\bar{b}}$. Also, for the pair of sites L",M", the curve a"b"c"d" gives the available outlay-time-delay combinations for both resource-bundles. It would seem, at first glance, that the factorability condition is satisfied here, with the ideal weights of the bundles being equal. Yet this is not necessarily so. Suppose time-costs are relatively low for $\bar{b}$, and relatively high for $\bar{\bar{b}}$; let UU be an iso-cost curve for $\bar{b}$, and VV and WW be iso-cost curves for $\bar{\bar{b}}$. (For simplicity, we let these be straight lines, as well as letting the available outlay-delay combinations be connected curves; the argument is not affected by this). For the transportation of $\bar{b}$, the combination b' is chosen for the trip from L' to M', and the combination b" is chosen for the trip from L" to M". These give the same total cost. For the bundle $\bar{\bar{b}}$, the combinations chosen are c' and c", reppectively, which give different total costs. Therefore, the metrics assigned by $\bar{b}$ and by $\bar{\bar{b}}$ are not in proportion, and factorability breaks down.

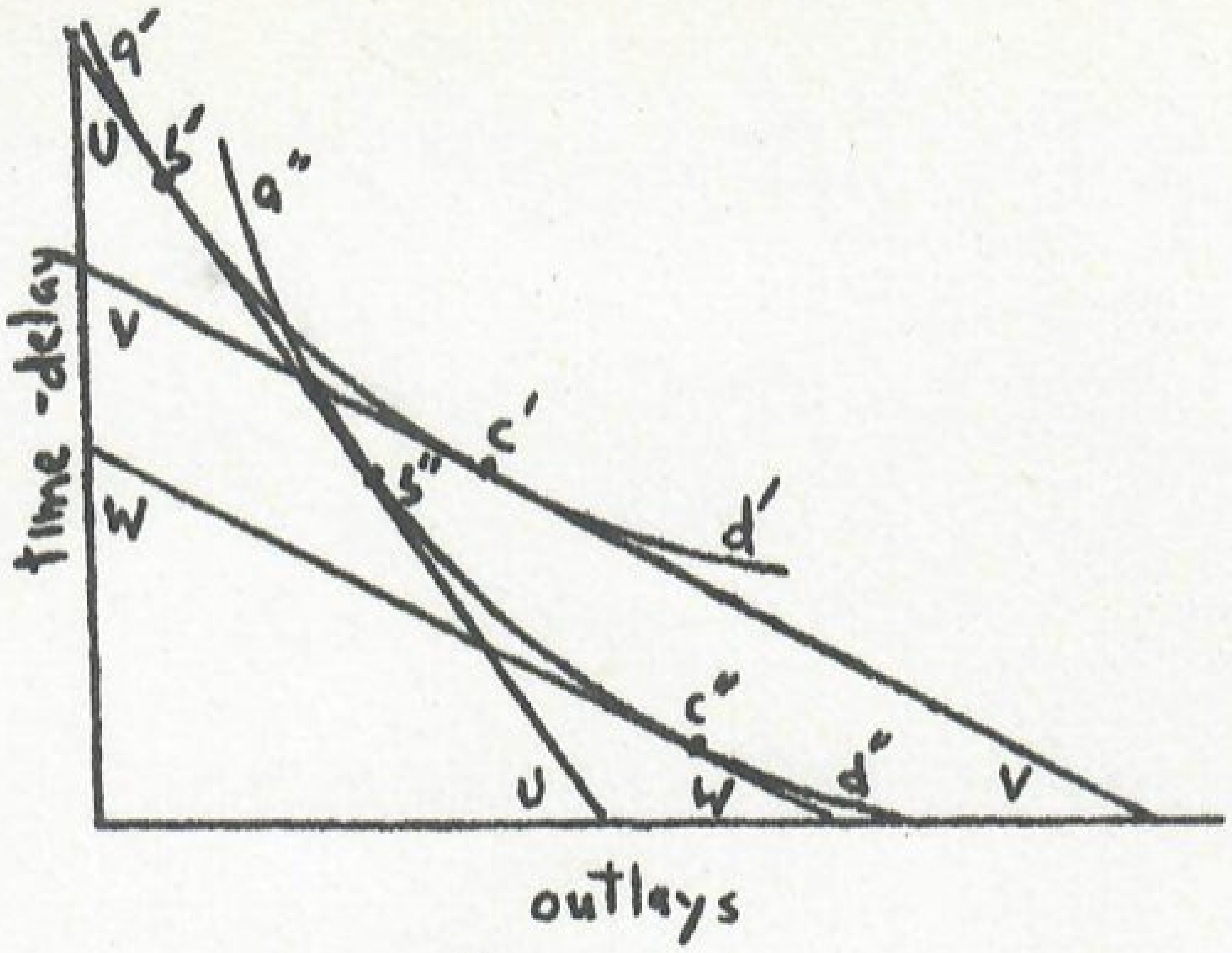

Figure 1

3) Discriminatory tolls, tariffs, and barriers to movement may destroy factorability.

None of the metrics which we have defined need bear any close resemblance to the metric defined by geographic distance.#

* By geographic distance between two points is meant the physical length of the shortest curve connecting them and lying on the surface of the Earth. If the Earth were a perfect sphere these would, of course, be great circle distances.

Nature and man combine to distort cost-metrics away from geographic metrics; irregularities in elevation, the distribution of land and water, wind and water speeds, texture of surface, climate, the distribution of road- and rail-nets, pipeline and electric grid systems, ports and airfields, being some of the contributing factors. On top of this are powerful institutional factors: common carrier routes, tariff systems, movement barriers (especially at national frontiers), perhaps bandits and pirates.*

* A graphic picture of the resulting jumble, which takes account only of the outlay component, may be found in J.W. Alexander, S.E. Brown, and R.E. Dahlberg "Freight Rates: Selected Aspects of Uniform and Nodal Regions" Economic Geography 34:1-18 January, 1958.

To round off the picture we should bring in the time-variability of metrics. One may distinguish daily, weekly, and annual cyclic components, and secular trends. Daily components arise through peak-load congestion, toll variations (e.g. in telephone rates), and visibility conditions; weekly from the variation between work-a-day and week-end travel patterns, and their resulting congestion shifts; annual from climatic and travel-pattern cycles. Secular trend arises, of course, from transportation construction, from growing traffic flow, from the introduction of new transportation systems, and from growing private vehicle ownership with rising incomes.

In summary, the idealization involved in going from the real world to a cost-metric is fairly heroic. Analytic difficulties make it hard to do much better.

(As with ideal vs. real metrics, we may contrast ideal with real weights. There is considerable distortion in weight ratios, even between different-sized bundles of the same resource: "bulk-economies" make the ideal weight of two tons of a commodity less than twice the ideal weight of one ton.)

One final point deserves clarification. The "heroic idealization" does not consist in the substitution of ideal weights and distances for their physical counterparts. On the contrary, only the ideal weights and distances are economically relevant, and we may simply forget about physical weights and distances. It consists in (1) the compression of costs into a single number; (2) the assumption of factorability; (3) the attribution of a simple structure to the ideal metric (e.g. the Euclidean plane assumption, as in a later chapter).

2.3. Flow-Patterns and Market Areas

In this section we deal with just a single commodity. There are a finite number of sites. For any pair of sites L, M, we are concerned with just the qualitative attribute of the existence or non-existence of a flow of the commodity from L to M.#

# This flow-pattern can be represented as a directed graph, an arc running from L to M if and only if there is a flow from L to M. All the following definitions thus have immediate graph-theoretic counterparts.

The sites may be partitioned into four categories. A source point is a site having outflows to other points but no inflows from other points. A sink point is a site having inflows from other points but no outflows to other points. An isolated point is one having no flow connections with any other point. All remaining sites are called trans-shipment points. Among trans-shipment points may be distinguished assembly points, having inflows from several points and an outflow to just one; break-of-bulk points, having outflows to several points and an inflow from just one, and through-points, having an inflow from just one point and an outflow to just one point.

A flow pattern is acyclic iff there is no cyclic route of different sites such that there is a flow from each site to its immediate successor (roughly, iff there is no sequence of flows going around in a circle). A flow pattern is one-many iff no site has an inflow from more than one point. It is many-one iff no site has an outflow to more than one point.

A supply system is a flow pattern which is acyclic and many-one. A demand system is a flow pattern which is acyclic and one-many. Site K is linked to site M iff there is a sequence of sites, $L_1, \ldots, L_N$, (possibly empty), such that there is a flow from K to L_1, from L_1 to $L_2, \ldots,$ from L_{N-1} to L_N, and from L_N to M. The supply area of a sink point is the set of all sites which are linked to it. The demand area of a source point is the set of all sites to which it is linked. It is convenient to include the sink point itself in its own supply area, and the source point itself in its own demand area. Market area is the generic term for either a supply or a demand area.*

* This terminology for "supply", "demand", and "market" areas follows Lösch, and brings out the basic symmetries better than the usual terminology, which calls our demand areas "market areas", and has no generic term. Cf. A. Lösch The Economics of Location (W.H. Woglom, translator)(New Haven, 1954) page 9; E.M. Hoover The Location of Economic Activity (New York, 1948) Chap.4; W. Isard Location and Space- Economy, op. cit., Chap.7, Secs.1,2.

It is easy to prove the following basic result.

Theorem: Every site in a supply system is either isolated, or belongs to exactly one supply area; the number of supply areas equals the number of sink points. Similarly, every site in a demand system is either isolated, or belongs to exactly one demand area, the number of demand areas equalling the number of source points.

Figure 2 depicts a demand system, an arrow indicating a flow. It has 18 sites, 3 demand areas and source points, 3 isolated points, 9 sink points and 3 break-of-bulk points.*

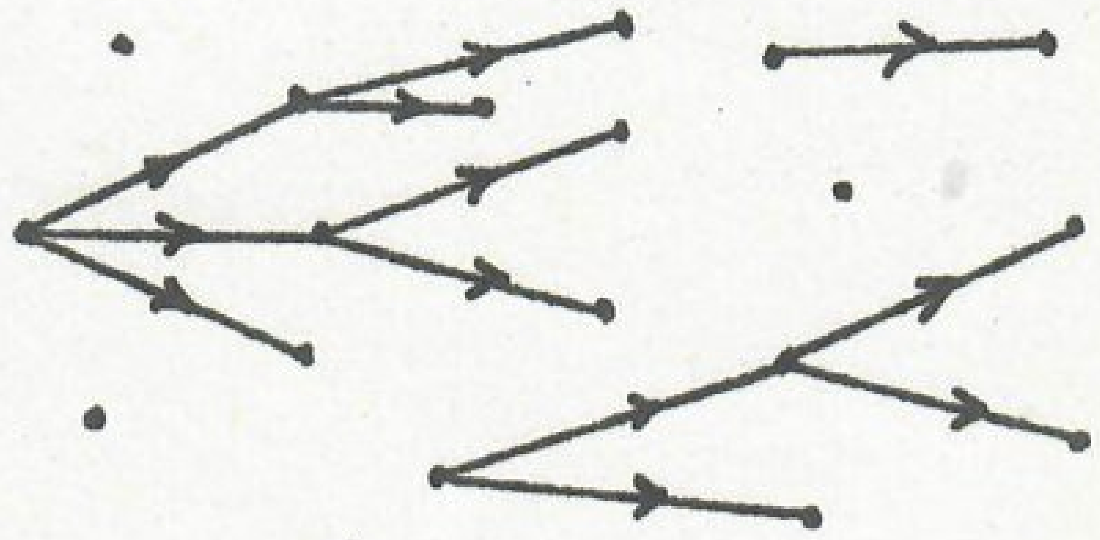

Figure 2

* The one-many acyclic relations, also known as hierarchies, play a fundamental role, not only here, but in the theory of organizations, the theory of games, genetics, taxonomy, and many other places.

2.4. Metrics, Flow-Patterns, and Price-Fields

We will now develop some of the interrelations between the metrics of 2.1 and the flow-patterns of 2.3. To do this we introduce still a third spatial system, the price field, which is simply a function assigning a real number, the local price of the commodity, to each site.

In an interpretation, the local price is simply identified with the actual commodity price at a site (if it exists), with the understanding that the bundle which is being priced is the same as the bundle defining the metric in 2.1. However, actual prices will in general exist only at the subset of sites at which exchanges of the commodity occur. There are a number of ways in which the gaps in the price field may be filled. For our purposes we choose the simplest, which is to assume that there is a way of assigning real numbers to the sites at which no exchanges occur, such that the resulting full price field satisfies the relations postulated below. These numbers may be called virtual local prices.

We are continuing here with our static approach, so that no account is taken of price fluctuations, or the stopping, starting, or even reversing of flows over time. As was mentioned above, the whole approach may be "dynamized" by interpreting sites not as points of space, but of space-time, so that prices are spread out over time as well as space, and flows move from an origin and time of departure to a destination and time of arrival. This line will not be pursued.

We now postulate the following connections among price-fields, flow-patterns and metrics:

1) $P(L) + d(L, M) \geqq P(M)$ for all L, M; and,

2) if there is a flow from L to M, $P(L) + d(L, M) = P(M)$, where p(L) is the local price at site L.

These are well-known relations. They may be interpreted either as competitive equilibrium conditions, or as normative conditions for efficient flow-patterns. In the first interpretation, (1) states that the activities: buying at L, shipping from L to M, and selling at M does not yield positive profits in equilibrium for any pair of sites; (2) adds that, if a flow occurs, profits are not negative either. In the second interpretation, the p's are "shadow prices", and (1) and (2) are dual relations to various minimal cost flow problems.*

* See P.A. Samuelson "Spatial Price Equilibrium and Linear Programming" American Economic Review 42:283-303 June, 1952; L.R. Ford, Jr., and D.R. Fulkerson Flows in Networks, op.cit., Chapter III.

We now examine some simple consequences of (1) and (2).

Theorem 1: If K is linked to M, then $p(K) + d(K,M) = p(M)$.

Proof: Suppose K is linked to M; then there is a sequence of sites K, $L_1, \ldots, L_N$, M such that there is a flow from each site to its immediate successor. We now apply (2) several times over:

$p(K) + d(K, L_1) = p(L_1)$;

$P(L_1) + d(L_1, L_2) = p(L_2)$;

- - - - -

$P(L_N) + d(L_N, M) = P(M)$; adding and simplifying, we get

$P(K) + d(K, L_1) + d(L_1, L_2) + \ldots + d(L_N, M) = P(M)$.

From the Triangle inequality, $d(K,M) \leq d(K,L_1) + d(L_1,L_2) + \ldots + d(L_N,M)$. From these last two results, we get

$$P(K) + d(K,M) \leq P(M) \; ; \quad \text{but} \quad P(K) + d(K,M) \geq P(M),$$

from (1). $\therefore \; P(K) + d(K,M) = P(M)$ QED

Theorem 2: If there is a flow from K to L_1, from L_1 to L_2,..., from L_{N-1} to L_N, and from L_N to M, then K, L_1,..., L_N, M is a geodesic.

Proof: Proceeding along the lines of the last proof, we find that $P(K) + d(K,L_1) + \ldots + d(L_N,M) = P(M)$;
but also, by Theorem 1, $P(K) + d(K,M) = P(M)$;

$$\therefore \; d(K,M) = d(K,L_1) + d(L_1,L_2) + \ldots + d(L_N,M).$$ QED

Theorem 3: Flow-patterns are acyclic.

Proof: This follows immediately from Theorem 2, and Theorem 3 of 2.1. QED

Corollary: For no pair of distinct sites, L and M, can there be flows both from L to M and from M to L.

This is a special case of cross-hauling. In common usage, "cross-hauling" refers to a situation in which there are two pairs of sites, (J,K) and (L,M), with J and K being relatively close, L and M being relatively close, the pairs being relatively distant from each other; yet J ships to M and L ships to K (see Figure 3). Let us formalize this.

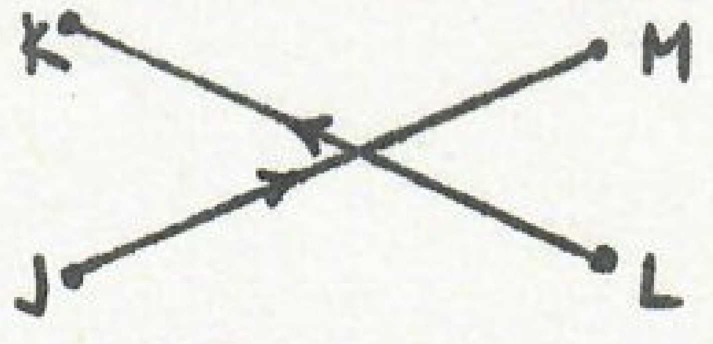

Figure 3

Cross-hauling is said to occur when there is a flow from J to M, a flow from L to K, and $d(J,K) + d(L,M) < d(J,M) + d(L,K)$.

Theorem 4: Cross-hauling does not occur.

Proof: The flows imply that $p(J) + d(J,M) = p(M)$, and $p(L) + d(L,K) = p(K)$, by Postulate (2); also $p(K) \leq d(J,K) + p(J)$, and $p(M) \leq d(L,M) + p(L)$, by Postulate (1); adding these four relations, and simplifying, we obtain $d(J,M) + d(L,K) \leq d(J,K) + d(L,M)$, which contradicts the third part of the definition. QED

Suppose the flow-pattern is a demand system. Every non-isolated point is then in exactly one demand area. If Postulates (1) and (2) hold, we have the following simple result.

Theorem 5: Site L is in the demand area of a source point J which minimizes $p(j) + d(j,L)$ over all source points j.

Proof: Since the source point J' of the market area in which L sits is linked to L, we have, from Theorem 1, $p(J') + d(J',L) = p(L)$. On the other hand, for any other source point j, $p(j) + d(j,L) \geq p(L)$, by Postulate (1); therefore, J' minimizes $p(j) + d(j,L)$ over source points. QED

The intuitive meaning of Theorem 5 is, of course, that each site purchases from a source minimizing the sum of local price at the source plus transportation costs, if it purchases at all. (The minimizing source need not be unique, and in this case, the price-field and the metric are not sufficient to determine in which demand area the site will fall. This phenomenon occurs with a continuum of sites along the borderline between two demand areas. The difficulty introduced is minor.)

For supply systems, the result analogous to Theorem 5 is

Theorem 6: Site L is in the supply area of a sink point J which maximizes $p(j) - d(L,j)$ over all sink points j.

The proof follows the same pattern as that of Theorem 5.

Theorem 6 means, intuitively, that a site sells to that sink offering the highest price net of transportation costs to the sink.

Theorems 5 and 6 suggest an alternative concept of market area which has the property of incorporating all sites, including the isolated points. Suppose we have a demand system which satisfies Postulates (1) and (2). The potential demand area of source point J is the set of all sites L for which J minimizes $p(j) + d(j,L)$ over all source points j. That is, the potential demand area consists of those sites which find J the cheapest source of supply, counting in transportation costs.

The problem again arises of those sites for which two or more sources are equally cheapest. One may simply assume that there are no such sites. A more palatable approach uses the concepts of the next seation, and assumes that the set of sites for which sources are not uniquely assigned is a set of "measure zero", and we agree to ignore anomalies which occur only on such sets.

To contrast with "potential demand area", our previous definition, based on actual flows, will be referred to as effective demand area. With the understanding of the previous paragraph, which circumvents the non-uniqueness problem, we then have the result:

Theorem 7: In a demand system, the effective demand area of a source is a subset of its potential demand area.

Proof: By Theorem 5, all sites in the effective demand area satisfy the criterion for being in the potential demand area. QED

A completely parallel construction may be made for supply areas. In a supply system, the potential supply area of a sink point J is the set of all sites L for which J maximizes $p(j) - d(L,j)$ over all sink points j. The effective supply area is, as before, the set of sites which are linked to J. We then have the result:

Theorem 8: In a supply system, the effective supply area of a sink is a subset of its potential supply area.

This follows from Theorem 6, just as Theorem 7 followed from Theorem 5.

2.5. Measures and Access-Perspectives

A region is simply a collection of sites. Suppose we have in turn a collection of regions with the following property. If R' and R" are both in the collection of regions, then their union and their difference are both in the collection. That is, the region consisting of all sites belonging either to R' or to R" (or both) is in the collection, and the region consisting of all sites belonging to R' but not to R" is also in the collection. A measure is a function which assigns a real number to each region of such a collection, and has the additional property that, if R' and R" are two disjoint regions of the collection (i.e. they have no sites in common), then the measure of their union is the sum of the measures of each: $\mu(R' \cup R'') = \mu(R') + \mu(R'')$.*

* This definition is a good deal weaker than the usual mathematical definition of "measure", but it is adequate for our purposes.

Examples of measures abound. Let us take as our regions various parts of the Earth's surface. Surface area is a measure; so is residential population (at some given moment); so is value added (in some given time-interval). As a matter of fact, the great bulk of published statistics appear to be, either measures themselves, or simple derivations from measures (as, e.g., per-capita income data are derived from the measures total income and total population).

An important class of regions are defined by metrics. Let K be a fixed site, and r a non-negative number. The closed out-sphere of radius r about K is the region consisting of all sites L such that $d(K,L) \leqq r$; i.e., the set of all sites not further than distance r from site K. The closed in-sphere, similarly, is the region consisting of all sites L such that $d(L,K) \leqq r$. In our cost interpretation of the metric, the out-sphere is the set of all sites reachable from K at a cost of no more than r (e.g. outlay, or time-delay, etc); the in-sphere is the set of all sites from which K is reachable at a cost of no more than r.

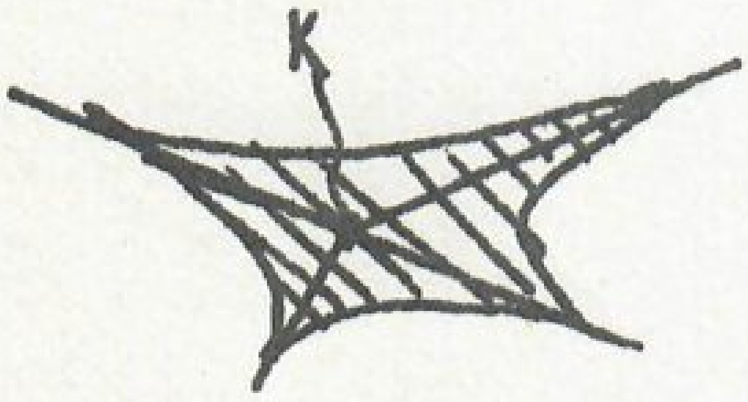

Figure 4

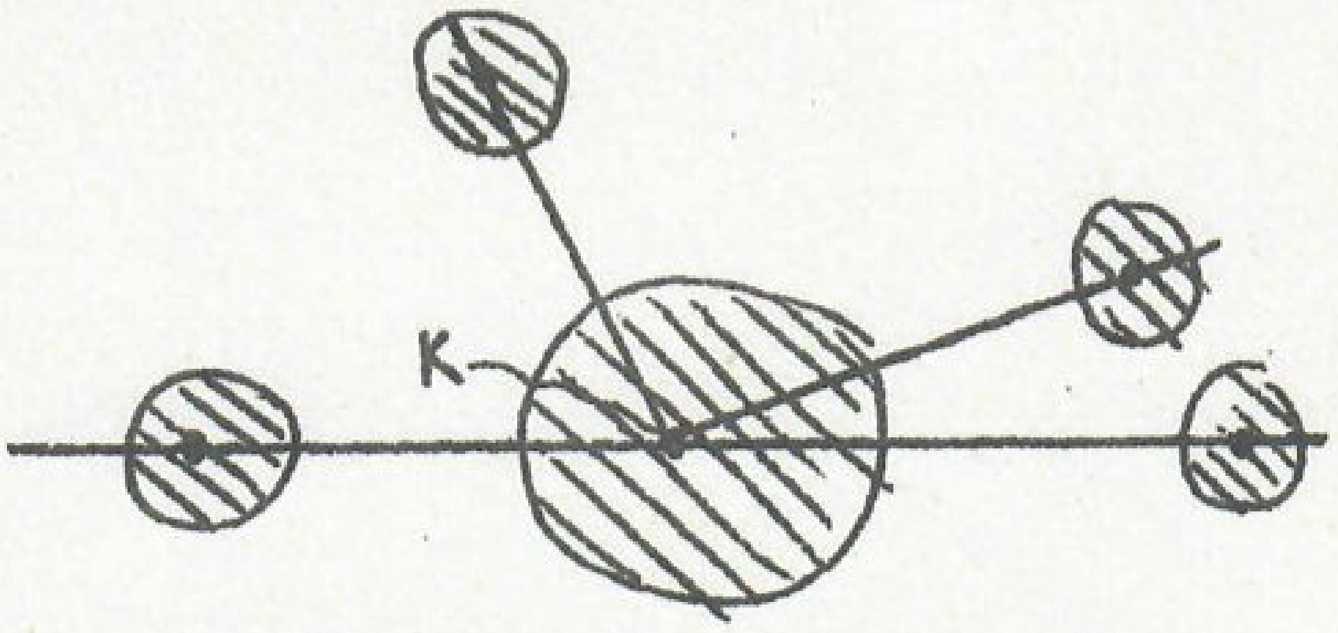

Figure 5

We should not expect these "spheres" to look very spherical in the geographical sense; (more exactly, we should not expect them to be circular discs lying on the surface of the Earth, as they would be approximately if cost-distance coincided with geographical distance.) Two important cases are shown in figures 4 and 5. In Figure 4, K is the hub of a radial system of transportation arteries. The spheres about K tend to extend "arms" along the arteries and have the typical starfish shape shown.

The situation in Figure 5 is similar, except that we are now dealing, e.g., with limited-access highways or commuter railways, so that one can get on or off the artery only at the dots. Here the "sphere"turns out to be not even connected, since an isolated circular disc around a dot is more accessible to K than some points further away, even if these points are geographically closer to K.

Now let us suppose we have a measure defined on all the in- and out-spheres about a certain site K; (it will also, of course, be defined on other regions besides these). The in-spheres about K constitute a one parameter family of regions, indexed by their radii, r. The <u>in-access perspective</u> of the point K is the real-valued function of a real variable, $\mu_K^{in}(r)$, $0 \leqq r < \infty$, which to every non-negative value of the independent variable r assigns the measure of the closed in-sphere of radius r about K. Substitution of "out-" for "in-" and repetition of the same construction defines the <u>out-access perspective</u> of the point K, $\mu_K^{out}(r)$.*

There are, of course, many access perspectives for a given site K: in general, a different one for every different combination of metric and measure. If the measure is non-negative--as is usually the case--the access perspective will be a non-decreasing function. This follows from the fact that a sphere of

* The access perspective seems to arise naturally as a way of describing the general accessibility of places. For example, in local booster advertising one often comes across statements of the form "30 million people live within a fifty mile radius of X-ville", which is just a specification of one value of the access perspective of X-ville, the measure being population.

smaller radius about K is a subset of a sphere of larger radius about K. If the former had a larger measure than the latter, the "ring" which is their difference would have a negative measure, contrary to assumption. The value of $\mu_K(0)$ (in- or out- alike) is the measure of the single site K itself.

As an example, suppose we confine ourselves to regions which are parts of the Earth's surface, and take surface area as our measure. For simplicity, let our metric be ordinary geographic distance. Since this is symmetric (i.e. $d(L,M) = d(M,L)$), in-access perspective and out-access perspective coincide for all sites. If we neglect the curvature of the Earth and assume we are on a Euclidean plane, we get, of course, $\mu_K(r) = \pi r^2$, the relation between the radius of a circle and its area. (This approximation is sufficiently good in many cases to justify its use as an assumption. A large fraction of the literature of spatial economics does so use it--the "homogeneous plain" assumption--and its simplicity compared with alternatives makes it likely that this will continue to be the case). More generally, if the Earth were a perfect sphere of radius ρ, access perspective would be $\mu_K(r) = 2\pi\rho^2\left(1 - \cos\frac{r}{\rho}\right) = \pi r^2 - \frac{\pi r^4}{12\rho^2} + \ldots,$ for $0 \leq r \leq \pi\rho$; $= 4\pi\rho^2$, for $\pi\rho \leq r$.

Of greater economic interest would be such measures as usable area or vacant area. These would incorporate factors such as geographical suitability, previous construction, and multiple-story floor space availability, and would, in general, have an irregular access perspective.

With the aid of access perspective, a large class of "intensity" or "accessibility" formulas can be constructed.

The *f-intensity at the site K*, with respect to a given metric and measure, is

3) $$\int_0^\infty f(r)\, d\mu_K(r),$$

where f(r) is a real-valued, positive and decreasing function, and the expression is a Stieltjes integral (which reduces to an ordinary Riemann integral is access perspective is differentiable). One may, of course, distinguish in- and out-f-intensities, defined by in- and out-access perspectives, respectively.

The f-intensity assigns a single number to the site K, rather than an entire function. If μ is non-negative, then f-intensity will be non-negative. Of the possible functions f(r) only one so far has attained popularity. John Q. Stewart introduced (3) with f(r) = 1/r under the name "potential"; (thus, "population potential", "income potential", etc., depending on the measure; the metric is almost always taken to be geographical distance).# No one has given a reason why this function--or any other, for that matter--should be favored, apart from correlations with other spatial distributions; (these correlations are, in general, much less impressive than the results for gravity formulas). The subject needs a theoretical grounding, perhaps along the lines of our crude attempt for gravity models in 1.7.

For a review, see W. Isard *Methods of Regional Analysis*, op. cit., Chapter 11.

3. <u>The Location of Weberian Activities</u>

In this chapter we pick up the thread of Chapter 1, concentrating on a different (and more traditional) set of spatial variables: what activities get carried on, and where? We start again with the individual decision-maker, then go on to study the equilibria resulting from the interaction of these decision. We also take up some normative issues, discussing the problem of optimal location patterns.

3.1. <u>The Individual Location Problem</u>

In Chapter 1 we discussed the individual's choice of itinerary. Here we discuss his land use decisions. These may be broken down into two phases: (1) what part of the Earth's surface does he acquire control over, and (2) to what uses does he devote this part?

Let us first discuss phase (1). The formulation is unusual; a more traditional approach is simply to ask, "where does the individual locate?" (We are using the term "individual" here to stand for any decision-making unit, whether it be an individual proper, a family, a corporation, a government body, or other organization). But this question misses three aspects of the location decision. First of all, one may choose a <u>number</u> of differenc locations separated from each other:(e.g. a firm may have offices, plants, and warehouses scattered throughout the world). Secondly, at each of these locations one must decide on the <u>size</u> of the parcel to acquire. Finally, at each location one must decide on the <u>shape</u> of the parcel as well.*

* Two noteworthy recent books in the small literature on parcel sizes are L. Wingo, Jr., <u>Transportation and Urban Land</u> (Washington;

Resources For The Future, Inc., 1961) and W. Alonso Location and Land Use (Cambridge, Harvard University Press, 1964). On parcel shapes, one pioneering approach is found in Alonso's book, Appendix B. There is also some work in the geographical literature on settlement morphology; cf. C.P. Barnes "Economics of the Long-lot Farm" Geographical Review 25:298-301 1935; M.Chisholm Rural Settlement and Land Use (London, Hutchinson, 1962). There appears to be no systematic work on the "numbers" problem, unless one includes the problem of optimal spacing in that designation.

All of these aspects are automatically contained in the problem of choosing a part of the Earth's surface. If we think of the surface as a collection of a large number of sites (perhaps a continuum of sites), then the problem is to choose a certain subset of these sites.

Not all conceivable subsets are possible options for our individual. We may divide constraints into institutional and budgetary. Under institutional constraints there are, in the first place, restrictions on minimal parcel sizes in the form of zoning and sub-division codes. (Older laws of primogeniture and entail perhaps belong in this category). In some places there are also maximal size restrictions, a resultant of land reform movements. Also, at any one time, a large fraction of the surface is not up for sale or lease at any price: much government land, perhaps places such as cemeteries or family estates which are held for sentimental reasons. (Some of these restrictions can be overridden if the individual holds the power of eminent domain--e.g. some government bodies and utilities). There may also be restrictions on some individuals as opposed to others, based usually on race, religion or nationality: e.g. segregation laws, alien settlement laws, restrictive covenents.

The budgetary constraint arises, of course, from the fact that one must pay a price to acquire controlof the remaining available parcels, and limited financial resources restricts most individuals to a very modest share of the Earth's surface.

Let us place the location problem in a more dynamic context. At any one time the Earth's surface is partitioned by ownership. (There remains a portion which is unclaimed by anyone--notably the high seas at the present time; there are also all sorts of difficulties involving vague boundaries and disputed territory; we ignore all these problems and assume that every site in the economy is assigned to one and only one owner). The pattern of ownership does not coincide completely with the pattern of control, since some sites will have been leased by the owners to other individuals who actually direct the land uses being carried on; also, some sites will be left standing vacant by their owners. Our individual finds the Earth's surface divided into four parts: land which he owns and controls, owns but does not control, controls but does not own, and land neither owned nor controlled; the middle two categories reverting to the control of their owners at various future times.

The natural units to work with in this dynamic context are not sites or regions perse, but space-time regions--i.e. spatial regions over an interval of time. Any one site is partitioned longitudinally into a succession of controllers (and also into a succession of owners, which need not coincide with the succession of controllers). One must decide where and when to sell or lease one's property, and conversely, where and when to buy or lease property from someone else, subject to the institutional and budgetary constraints. If the real estate market is imperfect, one has further decisions to make as to price demands or offers, and searching strategies.

In phase (2) of the problem--the use to which the land one controls is to be put--there are further constraints on action. Besides the obvoous restrictions imposed by technology and the general prohibition of illegal activities, we single out for special mention zoning laws, which have the peculiarity of being dependent of geographical location, and typically put restrictions on type of industrial activity (e.g. multiple-family residential, office, commercial, light or heavy manufacturing), on height and bulk of structures, and on density of occupancy.

The actual real estate market is evidently a rather complex affair. We now present a schematic version which is more tractable for analytical purposes. One big auction occurs, with all landowners on one side of the market and all potential renters on the other side. The same individual can, of course, function both as owner and as potential renter; it is convenient to separate these roles, just as one separates the roles of an individual who is both businessman and consumer.

The simplest reasonable behavioral assumption concerning landowners that can be made is that each site owned is rented to the highest bidder for that site. Here we must think of the owner himself, in his role of potential renter, putting in a bid along with everyone else. If all other potential renters underbid this "reservation price", the owner uses the land himself.

Suppose (1) there is perfect information in the market; (2) the bid which a potential renter will put in for a site does not depend on who owns the site. It is easy to see that the final distribution of sites among controllers (i.e. renters) is then independent of the initial distribution of sites among owners (provided compensation is made for any shifts in real estate wealth among owners, as determined by the bidding in the real estate market). This is an important simplification.

This simplicity may be destroyed if either side of the market discriminates. For owners, this means they act as if the bid were adjusted by a "discrimination coeffieient", depending on the owner and bidder, and the highest of the adjusted values is the one chosen. For potential renters, this means that the bids themselves, for a given site and bidder, depend on the owner.* A special case is the discrimination by an owner in favor of himself as renter; in Becker's terminology this may be called "auto-nepotism". It reflects the psychic value of working one's own land, etc.*

By our assumptions, the individual qua owner is in equilibrium when every site he owns has been rented out to the highest bidder, which may be himself, and where the bids may be adjusted for discrimination. For the potential renter, equilibrium is somewhat more complicated. We assume perfect information and

* See G.S. Becker The Economics of Discrimination (Chicago, University of Chicago Press, 1957), pp. 6-9.

* Obviously, one finds much more owner-used land than the minute amount that would occur if ownership and control of land were independently distributed. This may be partly explained by auto-nepotism, but also by ignorance and other market frictions.

ignore the complication that would arise if several equally highest bids are made for the same site. A given collection of sites is an equilibrium choice for control by a certain renter only if (1) it satisfies the institutional constraints; (2) the renter is willing and able to outbid all other potential renters' existing bids for all sites in the collection; (3) for all sites not in the collection, they are either not available, or the renter is not willing to outbid the highest existing bid on them; ("not willing" is taken to include the case where a site is so expensive that the renter's budgetary constraint would be violated if he rented it).

These are only necessary, not sufficient, conditions for renter equilibrium, and it may be that an entirely different collection of sites is preferred. This point brings us to one of the major complications of the whole subject. The amount one is willing to bid for a site depends on the location of the other sites one will have acquired. As a rule--but not invariably--the acquisition of a site enhances the value of nearby sites relatively to the value of distant sites to the potential renter. This occurs because the activities planned by the renter on his sites generally call for relatively heavy flows of traffic and messages among these sites: whether it be his own commuting, or the flow of goods in an integrated plant system, or the flow of messages between headquarters and field offices, etc.; thus, a tight site pattern tends to save on transport costs (and in fact enhances the attraction of more transportation-intensive plans). The limiting case of this tendency is the coalescence of sites into large connected parcels. There are many manifestations of this tendency. Thus, plants will buy up excess land in hopes of inducing "linked " plants to settle there.* Large parcels tend to be worth more per square foot than small ones.*

Too much should not be made of this agglomerative tendency, and one can easily find counteracting forces. For example, acquisition of a site suitable for residence usually diminishes the bids one will make for nearby residential sites. Facing a downward-sloping demand function in a region leads one to limit plant size there and build elsewhere.

Given his collection of sites, the renter then chooses his most preferred pattern of land uses, subject to an overall budgetary constraint, and to constraints imposed by technology, zoning laws, etc., on each site.

We have so far dealt with three systems of decisions which the individual must make: (1) choice of itinerary (in Chapter 1); (2) choice of location; and (3) choice of land uses. A fourth system deals with other forms of transportation and communication: shipment patterns, the acquisition and disposal of resources other than land, dispatching of agents, sending of messages, etc. This is quite a broad and complex subject in itself, brt it will suffice for our purposes to make some radically simplifying assumptions concerning it: We assume that the "efficiency" conditions, Postulates (1) and (2) of Section 2.4, are satisfied by

* J.R.P. Friedmann The Spatial Structure of Economic Development in the Tennessee Valley (Chicago, University of Chicago Press, 1955) p. 35, 42f.

* The premium is known as "plottage value". See A. M. Weimer and H. Hoyt Principles of Real Estate (4th ed., New York, Ronald, 1960) p. 285f.

our individual's resource flow choices. Inflow and outflow schedules at each controlled site are determined by the land use chosen there, and this fact, together with the assumptions, will suffice to determine the full pattern of resource flows in the problems we take up in this work.

These four systems are highly inter-related. In the first place, the range of options in each system depends upon the choices made in the others. As just mentioned, there are material-balance identities connecting resource flow patterns and land uses. Some activities may require personal participation, and the decision to engage in them therefore constrains one's itinerary. The choice in each system determines a net outlay or revenue, and these must jointly satisfy a global budget constraint.

The relation between one's commuting span,and the collection of sites one controls, deserves special mention. The two are by no means identical. Many, or most, of the sites that one visits are in fact not controlled by oneself: e.g. students at schools, patients at hospitals, patrons at restaurants, movies, or stores, employees at work; conversely, one may control activity at a site without visiting it: e.g., by telephone, or through agents. On the other hand, one is barred from admission to many uncontrolled sites of a restricted or private nature except in special circumstances;(e.g. one is admitted to the homes of friends, but may be barred from those of strangers; one is barred from certain military areas without a permit, etc.); in these cases one's commuting span is restricted by lack of control.

Secondly, one's preferences among the options in each system will, in general, depend upon the choices made in the others.

For example, the utility functions used in Chapter 1 to determine itineraries will, in general, contain parameters depending on the individual's choice of controlled sites, land uses, and resource flows. The itineraries themselves will, therefore, depend on these choices. Conversely, these choices will depend on the itinerary chosen, so that the full solution to the spatial problem for the individual can be attained only by satisfying a system of relations simultaneously.

(If the "individual" we are dealing with is actually an organization of some sort, containing N members, the itinerary problem of Chapter 1 broadens into the problem of determining N itineraries. All the old problems remain, and new ones appear. For example, utility may depend on who meets with whom, when, where, and for how long; more generally, utility depends in some non-trivial manner on all the itineraries, which introduces the problem of co-ordination. We will not pursue this line of inquiry in the present work.)

3.2. Weberian Activities

We begin by stating a basic location principle. Suppose an individual has decided all aspects of his spatial plan--all sites to be controlled, all land uses, commuting pattern, and resource flows--except for one particular: he has not decided whether to use site L' or L" for a certain activity. L' and L" are assumed to be identical in all respects save that of location: (1) technically, the sites are equally feasible for the contemplated activity; (2) they are both zoned to permit the contemplated activity; (3) both are available to the individual. Assume further that our individual does not discriminate, and that he has perfect information, and that he has no sentimental preference for one of the sites per se over the other (that is to say, he is indifferent to absolute geographical location). Finally, assume that outlay and time-delay are the only relevant aspects of transportation.

Substitution of the site L" for the site L' means the following: (1) the activity planned for L' takes place at L" instead; (2) all resource flows planned to go to L' go to L" instead, keeping their original points of departure; (3) all resource flows planned to emanate from L' emanate from L" instead, keeping their planned points of arrival; (4) sojourns at L" are substituted for sojourns at L'; (5) resources planned to be acquired or disposed of locally at L' are instead to be acquired or disposed of locally at L". Time schedules for arrivals, departures, purchases and sales at L" are to be the same as planned for L'.

What relevant differences remain between the two plans? There are possible differences in transportation outlays, since the plans differ in traffic flow pattern; there are possible differences in local prices at the two sites--in particular, rents may differ. In short, the only differences that remain are purely pecuniary.

The basis for choice between the two plans is now clear. One adds total transportation outlays to total net on-site outlays, including rent, under the two plans (properly discounting everywhere), and chooses the plan with the smaller overall total outlay. If one happens to own one or both of the sites, rental still remains as an opportunity cost, and one's final choice will not be affected.

The assumptions leading to this conclusion are rather severe. However, one should also note what assumptions need not be made about motivation. One does not have to assume profit maximizing behavior, or in fact anything more than the trivial supposition that, other things being equal, more money is preferred to less.

If there are several contending sites instead of just two, the same reasoning leads to the conclusion that one chooses the site minimizing the sum of (discounted) transportation outlays plus net on-site outlays, including rent. This will be called the <u>site-substitution principle</u>.*

One broadening of the site-substitution principle that suggests itself is to go from transportation outlays to transportation costs, including time costs. That is, we relax the very restrictive assumption of identical time-schedules for the contending plans, and let outlays and time-delays depend on sites of origin and destination, as discussed in previous chapters. One then converts these time costs to dollar equivalents, and minimizes the grand total. Very likely this would have to be done in any practical application of the site-substitution principle. But it raises several thorny problems. Revised schedules of arrivals and departures at a site may turn out to be technically infeasible. Even if this is not so, one will, in general, not be indifferent between the original activity and its revision. The problem of converting time to dollars to take account of this fact is formidable. Time, like space, is not a homogeneous commodity, but a continuum of different commodities.*

* This is a special case of the general principle of substitution, which is the central theme of Isard's <u>Location and Space-Economy</u>, op. cit. Chapters 5 and 6 deal with the present situation. It is not made quite clear that a profit maximization assumption is not necessary for the main results.

An activity may be represented (not necessarily uniquely) by (1) a non-specific site (i.e. a spatial region of a certain size and shape); (2) a non-specific time interval; (3) an initial bundle of resources; (4) a terminal bundle of resources; (5) a stream of resource inflows defined on the time interval; and (6) a stream of resource outflows. A special case is a steady-state activity, in which initial resource bundle is the same as final resource bundle, and the inflow and outflow streams do not depend on time.

The size and shape of the site may be used for a further classificatory breakdown. Consider, for example, (1) a multiple-story structure; (2) a farm; (3) a railroad; (4) a small manufacturing plant in a rural area. As far as the location of the plant is concerned, a single geographical point (given, say, by a longitude and a latitude) may be an adequate descriptive representation. The railroad may be adequately described by a network of arcs tracing out its rail pattern. For the farm we may have to specify the portion of the Earth's surface over which it spreads. Finally, an adequate analysis of multiple-storying requires explicit attention to the vertical dimension. We have used geometrical constructs of zero, one, two, and three dimensions, respectively, in going from the plant to the multiple-story structure. In reality, of course, all these activities occupy three-dimensional spatial realms. The point is, however, that in many cases we can and must simplify and idealize to focus on the essence of the problem in hand.

* cf. Becker "A Theory of the Allocation of Time", op. cit., and Section 2.2 above.

The very same activity may be represented adequately by a geometrical construct in one problem, and inadequately by the same construct in another. Suppose we take a textile mill in the suburbs of a Southern community. For the problem of explaining why the plant is located in the Southeastern United States, a zero-dimensional point description of its location is adequate. To explain its distance from the central city, on the other hand, its two-dimensional surface spread becomes relevant. Finally, to explain how it fits into the "skyline" distribution of the urban area, the mill must be considered as a three-dimensional entity.

The dimensional characteristics of activity locations enter into the analysis of problems in two ways, one related to metrics, the other to measures. The point differs from the line, surface, and volume, in being "simply-located". that is, given the metric, its distance to and from all other points is uniquely determined. For the other realms, a point of ingress or egress must be additionally specified in order to fix distances. Secondly, the point and the line, as opposed to the surface, are of zero areal spread. (As for the volume, we need only consider its surface cross-section in this regard). The functional significance of this fact is that rent is a negligibly small item in outlays for these activities, and has no significance in determining their locations. (The surface vs. volume distinction will be considered in a later chapter).

We may formulate criteria of adequacy in terms of these distinctions: An activity is adequately represented as having a point location if (1) it is so compact that the ambiguity in its distance measure to or from all other relevant sites is negligibly small for the problem in hand, and (2) it is so small that rent is a negligible item in determining its location for the problem in hand.* If simple location breaks down but rents are still negligible, a one-dimensional representation is adequate. (But if we want to explain the fact that, e.g., roads are built to swerve around areas of high land values, we must consider the width of the road to be non-zero; we get a system of noodles rather than a system of arcs).

An important mixed case occurs when rents are not negligible, and yet the activity locus is so compact it may be represented as simply located. (This occurs in some problems of farm location, of urban land uses, and in Thünen systems in general). The difficulties which arise thereby will be discussed in Chapter 4.

Activities which can be represented by point locations will be called Weberian activities.* The great simplicity of this representation commends its use wherever one can get away with it. The whole theory of commuting of Chapter 1, for example, implicitly took sites to be simply-located.

* For example, in the location of manufacturing activities, rents are thought to be a negligible factor in determining broad regional patterns, but not negligible for location within a city. Hoover Location Theory and the Shoe and Leather Industries, op. cit., p. 76, 107 note 17.

* after Alfred Weber, well-known for his doctrine of plant location, which uses this representation. See Alfred Weber's Theory of the Location of Industries, (C.J. Friedrich, ed.) (Chicago, University of Chicago Press, 1928) (original edition, 1909).

We now apply the site-substitution principle to the case of Weberian activities. That is, the one activity whose location is still undecided is a Weberian activity. We have, therefore, to find the optimal geographical point at which to site this activity. The simplest case--and the one usually assumed--is the one where every point in the system is a candidate for the site. (More generally, some points might be excluded, e.g., by zoning constraints). We are then to minimize the sum of transportation outlays plus net on-site outlays, including rent, over all the points of the system. By the Weberian assumption, rent is negligible at all points, so one term of the sum drops out. As a matter of fact, the interesting cases, so far as generality of results go, are those in which all net on-site outlays drop out or can be converted into transportation costs. The only costs left to vary from point to point are transportation costs.*

The way in which net on-site outlays may be converted into transportation costs (when this is possible) may be illustrated by Figure 1. We concentrate on some resource which is an input to our Weberian activity. Suppose that its flow pattern is a demand system, and let the two irregular regions in Figure 1 be demand areas, with source points S' and S" inside their respective areas. Let K, L, and M be three contem-

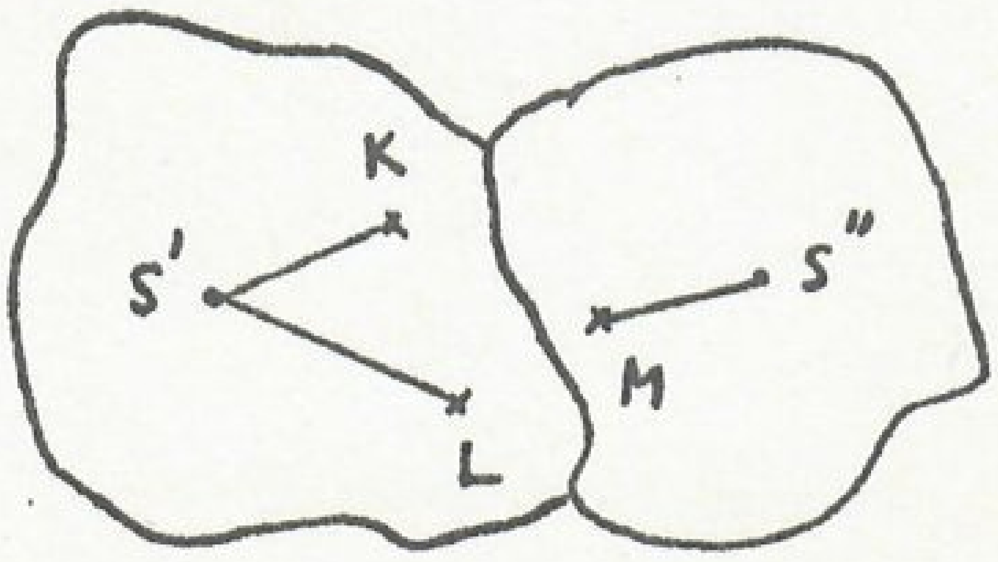

Figure 1

* This is the case of "transport orientation"; see Alfred Weber's Theory of the Location of Industries, op. cit., Chapter 3; Isard Location and Space-Economy, op. cit., Chapter 5; Hoover The Location of Economic Activity, op. cit., Chapter 3.

plated sites for our Weberian activity. Neither S' nor S" need be one of the other sites controlled by our individual. Let us assume first that neither one is controlled. (For example, our individual may buy from local stores at K, L, or M, transportation of the resource from its sources to local stores being carried out by other agents.) Sites Kand L are supplied by S'; site M is supplied by S".

By Postulate (2) of Section 2.4, $p(S')+d(S',K) = p(K)$, and $p(S')+d(S',L) = p(L)$. By subtraction, $d(S',K) - d(S',L) = p(K) - p(L)$. Replacement of the local prices $p(K)$ and $p(L)$ by the distances $d(S',K)$ and $d(S',L)$, respectively, would alter the total cost figures for K and L by the same constant, $-p(S')$, and so would leave their rankings unchanged and lead to the same choice of location. The same argument applies to all potential locations within the demand area of S'. If we apply the same procedure to site M, however, we find that replacement of $p(M)$ by $d(S",M)$ alters the total cost figure for M by $-p(S")$. If $p(S')=p(S")$, the replacement of local prices by distances preserves rankings both within and between the demand areas. If the source prices are unequal, rankings need not be maintained <u>between</u> demand areas (e.g. K vs. M), though they still will be maintained <u>within</u> any one demand area (e.g. K vs. L).

A complication arises if, say, S' is one of the sites controlled by our individual, and the resource movement out of S' to the Weberian site is part of his overall plan. If the Weberian activity were located outside the demand area of S'--e.g. at M--the planned shipments from S' to M would violate the efficiency conditions, Postulates (1) and (2) of 2.4. We will assume that plans are altered to conform with the efficiency conditions, even though this undermines the "ceteris paribus" construction on which the site-substitution principle rests.*

If we concentrate on a resource which is an output of the Weberian activity, and whose flow pattern is a supply system, we arrive at analogous conclusions. Figure 1 may now represent supply areas, S' and S" being sink points. The replacement of local prices by cost-distances to the appropriate sink point preserves total cost rankings within supply areas, but not necessarily between them (unless local prices at sinks are equal).

We assume that each input or output of the Weberian activity which is acquired or disposed of locally may be treated in the foregoing manner. Each resource partitions the economy into a collection of (potential) market areas. The market areas for one particular resource need have no special relation to those of another (though the channeling of transportation routes and trade centers make for broad similarities of pattern). All the resources that enter as inputs or outputs into the Weberian activity are then assumed to be of two types. The first partitions the economy into market areas, as above, so that the site with which the Weberian activty is linked by the resource varies with the location of the Weberian activity. Resources of the second type are tied to one fixed site. Commuting offers an instructive example. Trip frequencies to and from the Weberian site (wherever it is) and each other site in the economy are specified by the plan. Formally, the individual as a commuter acts as if he were

* A similar situation arises in international trade, when one has a choice of buying from an expensive compatriot or an inexpensive foreigner. Here the "individual" is the nation as a whole.

a large number of resources of the second type, one going to each site to which the commuter makes trips from the Weberian site (a kind of output of the Weberian activity), and one coming from each site from which the commuter makes trip to the Weberian site (a kind of input of the Weberian activity).

Resources of the second type are in fact a simple limiting case of resources of the first type, in which the entire economy is embraced by a single market area. They therefore offer no new problems.

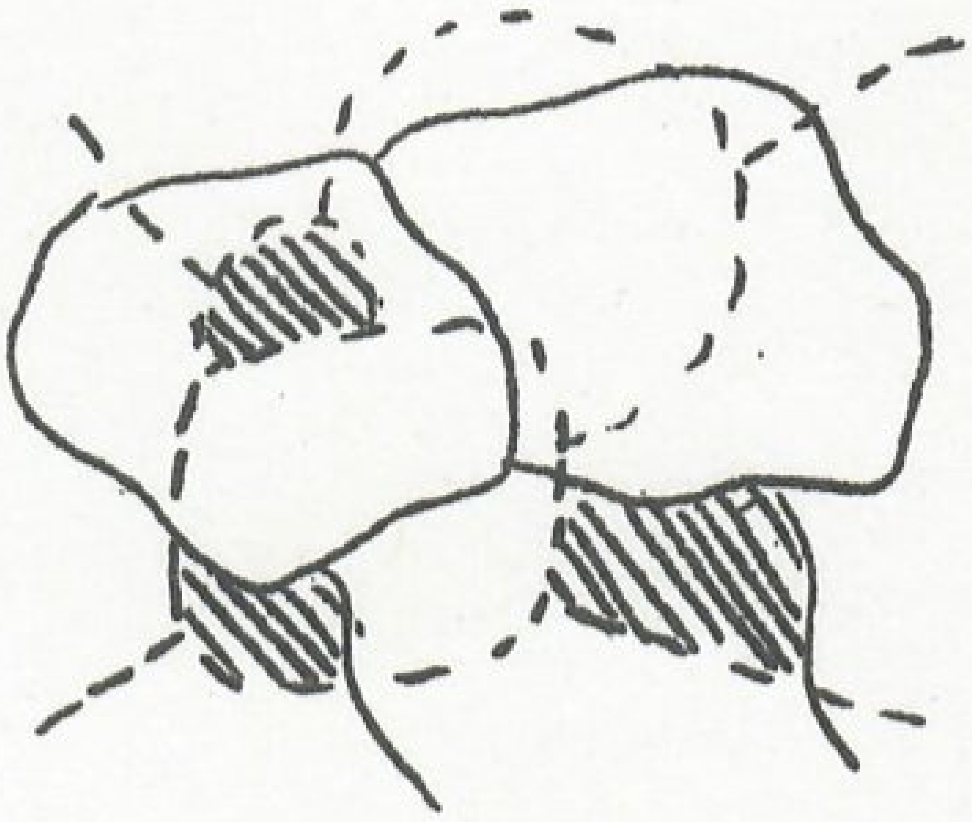

Figure 2

Let us now superimpose these market systems--one for each resource inputted or outputted from the Weberian activity. (Resources of type two may be omitted since they do not properly partition the economy.) Figure 2 depicts the superposition of part of two such systems, market area boundaries of one indicated by solid lines, boundaries of the other by dotted lines. An <u>elementary area</u> for a given set of resources is any non-empty collection of sites which is the intersection of market areas, one drawn from the market system of each of the resources in the set. For the set of two resources whose market areas are shown in Figure 2, the elementary areas are the smallest pieces into whichthe plane is divided by the two systems of borderlines combined; three of these are shaded. If all resources are of type two, there is just one elementary area: all sites combined.

The importance of elementary areas stems from the fact that a shift from one site to another <u>within</u> an elementary area of our Weberian activity does not cause a shift in the source point of any resource input or the sink point of any resource output; whereas a shift in the site of our Weberian activity between elementary areas causes a shift in the source or sink of at least one resource input or output; furthermore, one must worry about source and sink prices in the "between" case, while only transportation costs need be considered in making "within" comparisons. The upshot is that it is relatively easy to find the optimal Weberian location within each elementary area considered separately. The optimum optimorum can then be found by direct cost comparison of the "semi-finalist" candidate sites, one from each elementary area.

3.3. The Headquarter Location Problem

Suppose we are given an elementary area, and we are to find the optimal location within it for a Weberian activity, according to the site-substitution principle. We assume that the transport cost function is factorable, so that a single metric can be defined for all resources inputted and outputted from the activity. (Otherwise we might obtain as many different metrics as resources, and the problem would be most unwieldy). Ideal weights may now be assigned to all resource bundles.

The pattern of origins of resources which are inputs to the Weberian activity is fixed independently of the location of the latter within the elementary area; likewise, the pattern of destinations of outputs of the activity is fixed. Suppose a(L,t) is the resource bundle to be delivered to the Weberian activity site at time t from site L, and b(L,t) is the bundle to be delivered to site L from the Weberian site at time t. (For simplicity we assume time is discrete. Physically identical resource bundles moving at different times are to be regarded as different resource bundles, and may have different ideal weights; this enables us to allow for changes in transport costs over time, and for discounting. Discounting, for example, makes future bundles "lighter" than present bundles in ideal weight.)

Now consider a region R. Add up the ideal weights of all resource bundles a(L,t) over all $L \in R$ and over all times. This gives the total volume of resources to be moved from the region R to the site of the Weberian activity, measured in terms of the (discounted) total transport cost incurred per unit ideal distance. This will be called the <u>in-weight</u> of the region R. Analogously, addition of the ideal weights of all bundles b(L,t) over all $L \in R$ and over all times gives the <u>out-weight</u> of the region R. This is the total volume of resources to be moved from the site of the Weberian activity to the region R, measured in terms of the (discounted) total transport cost incurred per unit ideal distance.

We could work with in- and out-weights throughout the following analysis. However, for simplicity, we shall assume that ideal distances are symmetric (i.e. $d(L,M)=d(M,L)$). It is not hard to show that, with this symmetry assumption, the only relevant figure for each region is the sum of its in-weight and out-weight, which we simply call the <u>weight</u> of the region. The weight of region R is, then, the total volume of resources moving in both directions between the Weberian site and region R, measured in transport cost terms. Weight, in-weight, and out-weight are all non-negative measures.

Total transportation cost, which is to be minimized, is the sum, over all sites and times, of the product of ideal weight of resources moved, by ideal distance. When ideal distances are symmetric, this may be re-formulated in terms of regional weights:

Total transport cost when the Weberian activity is located at K is

$$1) \qquad \int r(K,L)\,d\mu(L),$$

where μ is the regional weight as a measure, and we have written r(K,L) in place of d(K,L) to avoid too many "d"'s.

For the situation considered in the last section there are just a finite number of sink and source points, and (1) reduces to a summation, one term for each such point. This case has been extensively treated in the literature, most elaborately when there are just three such points (forming a "location triangle"). We shall, however, deal with the general case, allowing for regional weights that are continuous, and mixed discrete-continuous. (For example, the case where a schedule of deliveries is to be made to a population distributed continuously over space will be covered).

The optimal site for the Weberian activity is the K which minimizes the integral (1) over all sites within the elementary area. Actually, we shall simply ignore this last constraint, and look for the minimizer over all possible sites. If the solution found happens to lie inside the elementary area then the original problem is solved. (This will always happen if the activity uses only type two resources, since there is just one big elementary area in this case).

We are thus led to the formal problem: Given the metric r(K,L) and the measure μ, minimize (1) over all sites K. (The metric is symmetric, and the measure is non-negative). If our space is the real line, with its customary metric, it is easily seen that the solution to this problem is the median of the distribution μ; (any median, if there is more than one). (The integral (1) is in this case proportional to the mean deviation about K, which is minimized at any median.) This property is sometimes used to define the "median" of more general distributions, and we shall follow this practice: The medians of the measure μ, with respect to the metric r, are the sites K which minimize (1).

For the next few paragraphs we specialize to the case of Euclidean spaces: The points of an N-dimensional Euclidean space are in 1 - 1 correspondence with N-tuples of real numbers such that $r(L',L") = \left[\sum_{i=1}^{N} (x_i' - x_i'')^2\right]^{\frac{1}{2}}$, where $x_1',\ldots,x_N'$ is the N-tuple corresponding to L', and $x_1'',\ldots,x_N''$ the N-tuple corresponding to L". The two-dimensional case is the one of greatest practical interest for spatial economics, but the one- and three-dimensional cases are also of interest (the latter especially since the advent of the "age of space".)

For such spaces, one may show that a median exists if measure is zero outside some sufficiently large sphere. Furthermore, if the entire mass of the distribution is not concentrated along a single straight line, one may show that the integral (1), when it exists, is a strictly convex function of position, which implies that the median is unique: that is, the occasional multiplicity of medians that one finds in one-dimensional distributions is a peculiarity of these alone. The integral, when it exists, is always a (weakly) convex function of position.

Point L" is called the reflection of point L' through the point K iff K is between L' and L" along a straight line, and r(K,L') = r(K,L"). Region R" is called the reflection of region R' through the point K iff the points of R" are the reflections through K of the points of R' (illustrated in Figure 3). A measure μ is

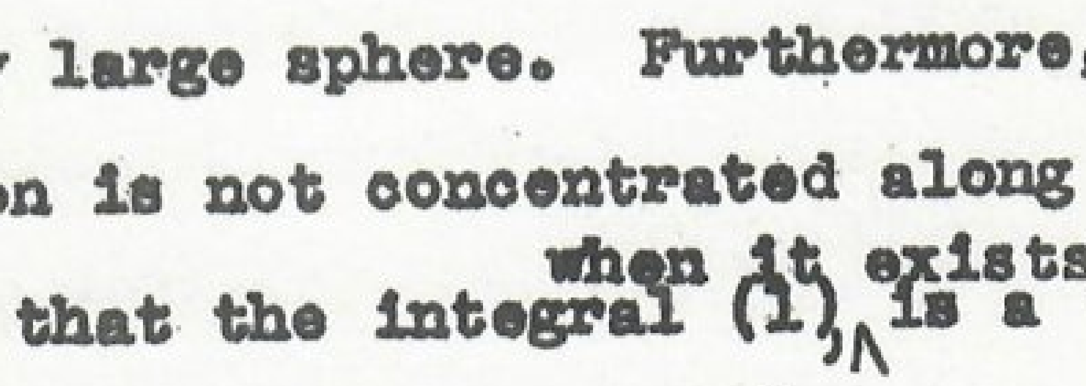

Figure 3

symmetric about the point K iff the measure of any region is equal to the measure of its reflection through the point K: $\mu(R')=\mu(R'')$ if R" is the reflection through K of R'. We have then

Theorem 1: If a measure is symmetric about point K in a Euclidean space, then point K is a median (if one exists).

Proof: Suppose point L' is a median, and let T(L') be the value of the transport-cost integral (1) at L'. Let L" be the reflection of the point L' through K. By symmetry, point L" must also be a median, so that $T(L')=T(L'')$. K is midway between L' and L", and the convexity of T therefore implies that

$\frac{1}{2}T(L') + \frac{1}{2}T(L'') \geqq T(K)$. Combining the last two results, we get $T(L') \geqq T(K)$. But T is minimized at L'; therefore, this must be an equality, and K is also a median. QED

Theorem 1 enables us to pinpoint the median in several simple cases, such as a measure uniformly distributed over a region bounded by a circle, ellipse, parallelogram, and even-sided regular polygon. Of these, the circle, square and regular hexagon are of most theoretical interest.

Point L" is called the reflection of point L' through the straight line S iff the line-segment $[L', L'']$ is perpendicular to, and bisected by, S. We may now define "reflection of a region through the line S" and "measure symmetric about the line S" in a manner analogous to the definitions for the point K, above. We then have:

Theorem 2: If a measure is symmetric about line S in a Euclidean space, then a median exists somewhere along line S (if it exists at all).

Proof: Suppose point L' is a median, and let L" be the reflection of L' through the line S. By symmetry, point L" must also be a median, so that T(L')=T(L"). Let K be the point midway between L' and L". K is on the line S, by definition of "reflection", and the reasoning in the proof of Theorem 1 shows that K is also a median. QED

Theorem 2 enables us to pinpoint the medians of uniformly distributed measures inside any regular polygon. (These have several lines of symmetry; these intersection of these lines must be the median, since it is on each, and unique). The most important odd-sided case is the equilateral triangle.

Analogous theorems may be derived for planes and hyper-planes of symmetry in higher-dimensional Euclidean spaces.

It may be objected that the results of the previous theorems are obvious; certainly no one will be startled by them. However, once one goes beyond these very elementary cases, the problem becomes extremely tedious to solve. (As exercises the skeptical reader is invited to determine where along their lines of symmetry the median lies for the measure uniformly distributed over a region bounded by (1) an isoceles triangle; (2) a semi-circle.) Furthermore, the topics treated in this and the next few sections are beset by "obvious" results which are very hard to prove, and which have sometimes turned out to be false. Under these circumstances it was thought useful to present these results, elementary as they are.

We now generalize the above median-location problem. Again we are presented with a metric and a non-negative measure; we are also given a positive integer N, and we are to find N sites,

$K_1,\ldots,K_N$, minimizing the following generalization of the integral (1):

$$2) \qquad \Sigma_{i=1}^{N} \int_{R_i} r(K_i, L)\, d\mu(L),$$

where R_1 is the region consisting of all points which are closer to point K_1 than to any of the other N-1 sites chosen. (Points which are equidistant from two or more of the K's may be allocated to the corresponding regions arbitrarily without affecting the value of the integral sum (2).) The median-location problem is just the special case when $N = 1$. The general case will be called <u>the headquarter location problem</u>.

This problem has a surprisingly large number of applications, and various special cases have been well worked over in the literature. If the sites $K_1,\ldots,K_N$ are thought of as sources (or sinks) for a certain resource, it will be noticed that the regions R_1 are the potential market areas corresponding to these sites in the special case when all local prices at the sources (or sinks) are equal. We shall refer to R_1 as the <u>service area</u> of the point K_1.

With the headquarter location problem, we have, in a sense, come around full-circle from our original point of view. We started with sources and sinks given, and tried to optimize the location of a Weberian activity which was linked to these. We then generalized from the case of a finite number of source and sink points to arbitrary measures; these measures were interpreted as volumes of traffic flows to and from each site in the space. The headquarter location problem now tmats the sources (or sinks) themselves as variables (with the restriction that local source or sink prices are everywhere equal).

The headquarter points $K_1,\ldots,K_N$ are still sites for Weberian activities, since they are taken to be simply-located and rent is ignored.

The following result gives a basic necessary condition for optimality in the headquarter location problem.

Theorem 3: Each headquarter point in an optimal solution to the headquarter location problem is a median of its own service area.

Proof: Suppose the statement is false. Then one can find a headquarter location problem with solution $K_1,\ldots,K_N$, such that, for some i, K_i is not the median of its service area R_i. There must, therefore, exist another point $\widetilde{K}_i$ such that

$$\int_{R_i} r(\widetilde{K}_i, L)\, d\mu(L) < \int_{R_i} r(K_i, L)\, d\mu(L);$$

$$\therefore \quad \Sigma_{\substack{j=1 \\ j\neq i}}^{N} \int_{R_j} r(K_j, L)\, d\mu(L) + \int_{R_i} r(\widetilde{K}_i, L)\, d\mu(L)$$

$$< \Sigma_{j=1}^{N} \int_{R_j} r(K_j, L)\, d\mu(L).$$

The right-hand side of this inequality is the total transport cost for the optimal solution. Now suppose we try the sequence of sites $K_1,\ldots K_{i-1}, \widetilde{K}_i, K_{i+1},\ldots,K_N$, identical to the original solution except that $\widetilde{K}_i$ has been substituted for K_i. Total transport cost for this collection of sites cannot exceed the left-hand side of this inequality, since the left-hand side represents total transport costs using the new sites and the old service areas, and service areas are chosen so as to minimize transportation costs given the sites. We have thus reached a contradiction, since a new solution has been found reducing the integral (2) below its alleged minimum.* QED

* A special case of Theorem 3 has been observed by Isard: Location and Space-Economy, op. cit., p. 233 note 24. Our proof follows his in all essential respects.

Theorem 3 may be generalized to

Theorem 4: Take any subset of sites in an optimal solution to the headquarter location problem, and take the union of the service areas associated with this subset. Suppose there are N' sites in the subset. Consider the headquarter location problem whose measure is identical to that of the original on the above union, and is zero elsewhere, and for which we are to place N' sites optimally. Then the subset must be an optimal solution to this new problem.

The proof of Theorem 4., which is entirely analogous to that of Theorem 3, is omitted. Theorem 3 is the special case $N' = 1$.

To illustrate the use of Theorem 3, we apply it to some simple one-dimensional problems. The solutions to these are well known. Suppose we have to place N headquarter points along a one-dimensional strip of length 1; the measure is uniform along the strip; the metric is Euclidean. Let $K_1, \ldots, K_N$ be the optimal placements, going from left to right (see Figure 4). The service areas are obviously intervals; let L_i be the border point between R_i and R_{i+1}, for i = 1, 2,.., N-1. We then have the following relations:

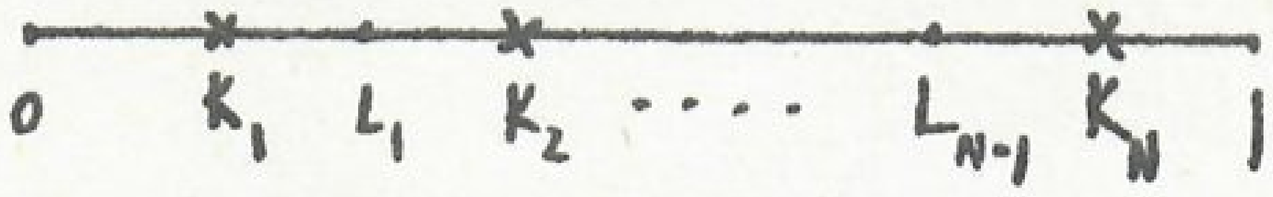

Figure 4

$$L_i = \frac{1}{2}K_i + \frac{1}{2}K_{i+1}, \quad \text{for } i=1,2,\cdots,N-1,$$

and $$K_i = \frac{1}{2}L_{i-1} + \frac{1}{2}L_i, \quad \text{for } i=1,2,\cdots,N.$$

In this second set of formulas we define L_0 as zero, and L_N as one. The first set of formulas follows from the definition of service area, since the point midway between K_i and K_{i+1} separates the points which are closer to K_i from the points closer to K_{i+1}. The second set of formulas follows from Theorem 3, since the median of a measure uniformly distributed over an interval is at its midpoint.

The solution to this system of relations is obvious at a glance: the successive points $0, K_1, L_1, K_2, \ldots, L_{N-1}, K_N, 1$ must be equidistant, from which it follows that $K_i = (2i-1)/2N$. This solves the problem of placing N headquarter points to minimize total transport costs to the nearest.

A variant of this is a loop (identify the points "0" and "1" in Figure 4). The same procedure as above gives the not unexpected result that equal spacing around the loop is optimal.

Theorem 3 gives only necessary, not sufficient, conditions for a placement to be an optimal solution to the headquarter location problem. This may be illustrated by a simple counter-example. Suppose we have a uniform distribution over a square, and headquarter points are placed in a row along a mid-line so that the service areas are congruent rectangular strips (Figure 5). Each point is the median of its service area;

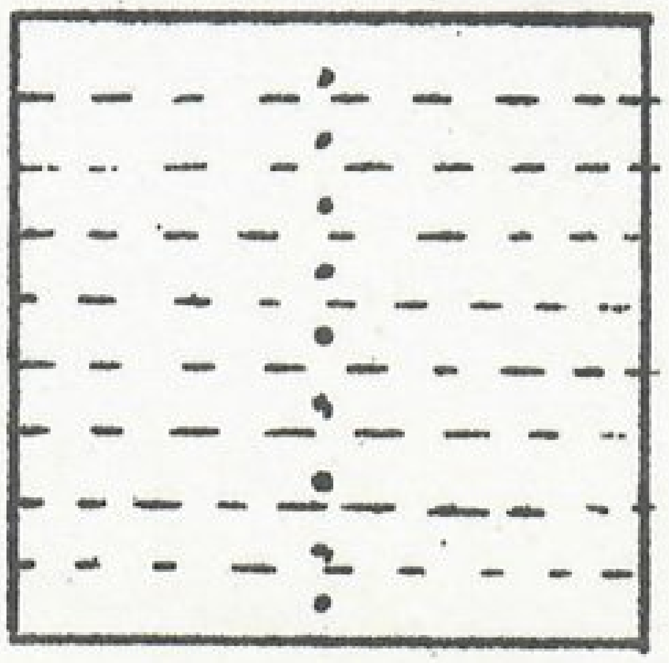
Figure 5

yet it is easily verified that a more diffuse scattering of headquarter points over the square lowers total transport costs, so the placement is non-optimal.

The proof of Theorem 3 suggests the following successive approximations procedure for solving the headquarter location problem.

(1) Pick N sites arbitrarily;

(2) Partition the system into service areas by assigning every point to the site closest to it out of the N picked in step (1);

(3) Find the median of each of these service areas: this gives N new points for a second approximation; using these repeat step (2); One then alternates betwwen steps (2) and (3) until a stable configuration is achieved.

It is easy to show that each successive round results in a set of N sites which have lower total transport costs than the last approximation (or at least no higher than the last). But it is not known under what conditions this procedure converges to an optimal solution. The above counter-example shows that such convergence is not universal.

Unlike the median, it need not be true, for distributions of dimension greater than one, that there is just one optimal headquarter placement. For example, take the problem of placing two headquarter points for a uniform distribution over a circular disc. By use of Theorems 2 and 3, it may be shown that the two points must lie along a diameter (see Figure 6). But obviously one diameter will do as well as another, so there are an infinite number of

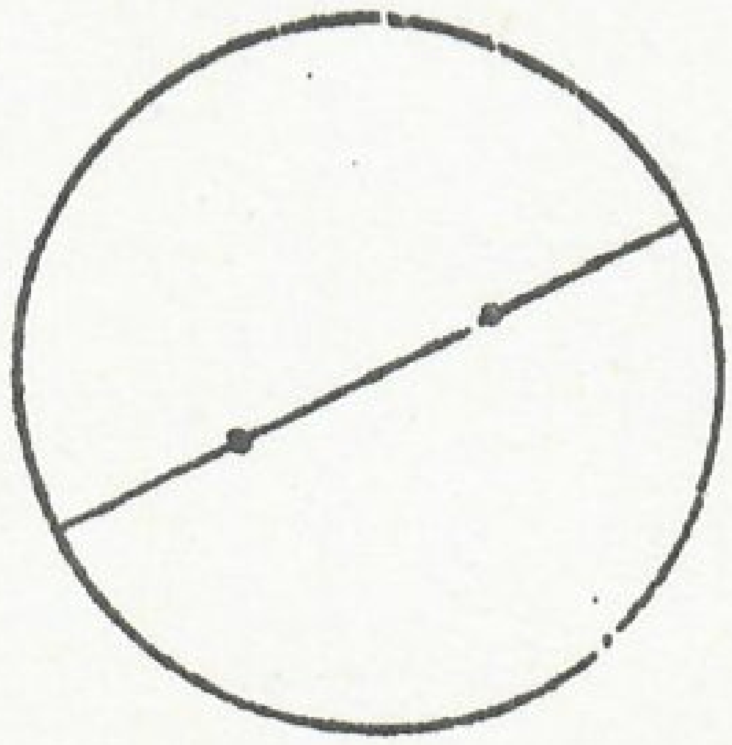

Figure 6

optimal solutions. Similarly, for a uniform distribution over the surface of an (ordinary) sphere, the two point problem is solved by any pair of antipodal points.

Now let us suppose we have a uniform distribution over an entire Euclidean plane. (This is the famous "homogeneous" or "featureless" plain of location theory, or at least an aspect of it). The headquarter location problem cannot be posed directly here, since total transport costs will be infinite in all cases. Let us, therefore, consider a sequence of problems which in some sense approach the plane as a limiting case. For example, take the problem: place N headquarter points optimally on a uniform distribution over a circular disc of area A (cf. Figure 6). We consider a sequence of such problems as both N and A go to infinity in such a way that the ratio A/N remains constant. An alternative approach would be to use the surface of a sphere instead of a circular disc, and let the sphere and numbers of headquarters expand to infinite in such a way that the surface area per headquarter point is constant.

We may now ask (1) does the placement pattern approach some stable form as N and A→∞? (2) if so, what is it? Contrary to a widely-held opinion, the answers to these questions are, strictly speaking, unknown at the present time. The usual assertion is that the optimal pattern is the hexagonal (or honeycomb) lattice shown in Figure 7. A honeycomb lattice is any one of a class of point sets. One of these sets consists of all points of the form $(1,0)m + (\frac{1}{2},\frac{1}{2}\sqrt{3})n$, where m and n range independently over the integers; all other members of the class may be obtained

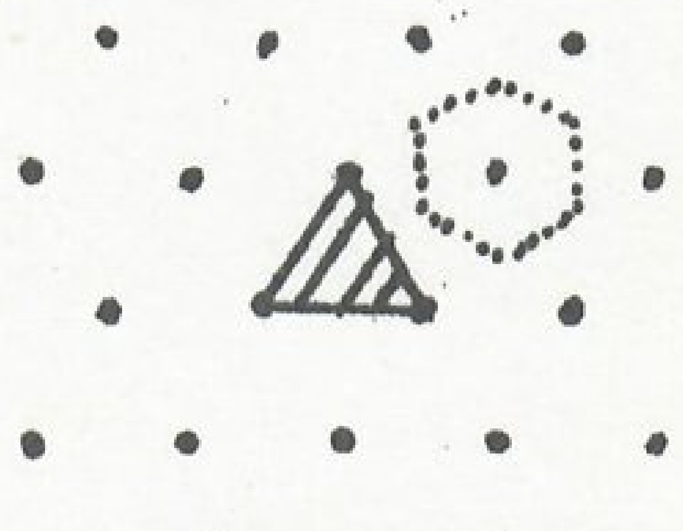

Figure 7

from this one by rotation, translation, and dilatation. If the plane were covered by a layer of equilateral triangles (one of which is shaded in Figure 7), their vertices would constitute a honeycomb lattice.*

We don't wish to deny this statement, which is probably correct. We do wish to point out that a genuine problem is involved, that the arguments adduced for this assertion are not sufficient, that a demonstration is probably quite difficult; and we shall offer a few suggestions for attacking the question.

The standard argument goes about as follows: First, the service areas are all polygons (convex polygons, in fact.) For reasons of symmetry one assumes that these polygons are congruent and regular. Since the service areas partition the plane, only three cases are possible: the service areas are either all hexagons, or squares, or equilateral triangles. One can show that the latter two cases are non-optimal.* This leaves the hexagonal

*cf. A. Lösch The Economics of Location, op. cit., p.110-114; W. Isard, Location and Space-Economy, op. cit., p. 240-242; E.S. Mills and M.R. Lav "A Model of Market Areas with Free Entry" Journal of Political Economy 72:278-288, June, 1964.
Actually these authors make an even stronger claim, that the honeycomb is optimal when demand depends in some general manner on price (in effect, the measure varies with the configuration of headquarter points chosen). The headquarter location problem for the homogeneous plain embraces only the special case of completely inelastic demands, in effect. The more general case will be touched on below, Section 3.6.

* A simple and ingenious sketch of a proof is found in Isard, loc. cit., p. 241f.

system reigning by default. The hexagonal service areas imply the honeycomb lattice pattern for headquarter points of Figure 7 (where one of the hexagonal service areas is dotted in).

The lacuna in this argument is the assumption of congruence and regularity of the service areas, a purely intuitive judgment.

A possible approach is to assume that the headquarter points form a general lattice, i.e., they are points of the form $(1,0)m+(a,b)n$, $(b\neq 0)$, where m and n range independently over the integers, or they are rotations, translations, or dilatations of this point set. (Geometrically, Figure 7 is "stretched" so that the little equilateral triangles become some general non-obtuse triangles, all still congruent to each other; the service areas are still congruent hexagons--though not regular, in general--or rectangles, when the little triangles become right triangles). One might then try to prove that the honeycomb lattice is optimal in this set. Our direct attempts have so far led to intractable integrals.

We conclude that, at the present time, the optimality of hexagonal lattices remains one of those obvious but elusive results we referred to earlier: a plausible conjecture, not a theorem.*

* For mathematical background on lattices, etc, see L. Fejes Tòth Lagerungen in der Ebene, auf der Kugel und im Raum (Berlin, Springer, 1953) and J.W.S. Cassels An Introduction to the Geometry of Numbers (Berlin, Springer, 1959).

3.4. Service Systems

In this section the headquarter location problem will be embellished in two ways. First, it will be embedded in the context of a more general problem; second, this problem will be given several concrete interpretations.

Suppose we are given a distribution of population over space. (This could be a population of individuals, families, firms of a certain kind, etc., depending on the interpretation). A set of headquarter points is to be found; at these, Weberian activities of a certain type are to be run. There is to be a flow of "services" between these headquarter points and the members of the population. The quantity of services going to any member of the population can be specified by a non-negative real number which is the ideal weight of the service. ("Services" may be flows of people, goods, or information, or some bundle of these, depending on the interpretation.) There is a metric, such that the total transport costs incurred by the flows of services is the product of ideal distance by ideal weight, summed over all flows. The total "production" of services at each headquarter point is equal to total flows of services from that headquarter point to the surrounding population. For each potential headquarter site there is a production cost function. Finally, for each member of the population there is a net benefit function depending on the level of flow of services to that member. This whole ensemble will be called a *service system*.

We are to find (1) the number and location of headquarter points; (2) the volume of flow between each headquarter point and each member of the population. By summation, the answer to (2) will also determine the level of "production" of services at each headquarter point. These will be the unknowns of the service system.

There are a large number of criteria by which these unknowns might be determined. In the next section, for example, we adopt a cost-benefit approach, and maximize: total benefits, summed over the population, minus total transport costs, summed over all service flows, minus total production costs, summed over all headquarter points. This approach is most appropriate for the case of public facilities, which are the most important class of interpretations for service systems. An alternative approach is that of a monopolist, who chooses the number, location, and production rates of his headquarter points (i.e. plants) to maximize profits. Still a third is an industry equilibrium approach, in which each plant is a separate firm, entry is free, and profits are maximized; this is the Löschian approach, and is discussed in section 3.6.

Embedded in this overall problem of service system design are several fragmentary problems, some already discussed, some new. Given the production pattern at headquarter points and the consumption pattern by the population, the detailing of origin-destination flows is a transportation problem. We will assume the operation of the efficiency postulates of section 2.4 in all cases; this will lead, in general, to the formation of demand areas with headquarter points as sources. Given total production at a headquarter point, its market area, and the distribution of population over this region by location and benefit function, there is the problem of optimal distribution among this population (e.g. to maximize benefits minus transport costs). Under certain uniformity assumptions, our original headquarter location problem will also turn up (see below).

We now come to interpretations. The most familiar, and the best-explored, is the industry location interpretation, in which the headquarter points are manufacturing plants which distribute a commodity over their respective demand areas. Closely related is the interpretation of headquarter points as warehouses and as retail stores. In all three cases we are dealing with the distribution of commodities. However, the conditions of transportation differ in that retail distribution generally entails a special "shopping trip" by the buyer; this fact, together with the relatively small size of package in retailing, means that the ideal weight of commodities sold at retail is high in comparison to the ideal weight of the same commodities sold at the manufacturer's or wholesale level; (and this in turn is a partial explanation of the abundance of retail outlets, as we shall see).

The distribution of water, gas and electricity, from pumping or power stations, introduces the new wrinkle of having special purpose transportation systems--pipelines and wires--which must be connected directly to the places where consumption occurs. It is not clear whether our simple transport cost assumptions are adequate for these systems. Television cable systems are of the same sort, with community antennas playing the role of headquarter points. Similar also are sewage disposal systems, leading to treatment plants as headquarter points--except of course that the "commodity" flow is inward, and we have a supply system instead of a demand system. This reversal raises no new difficulties at all: "consumer" benefits depend on the quantity of sewage removed, and at headquarter points we have disposal costs instead of production costs. Garbage collection is, in effect, a warehouse distribution system in reverse.

A great variety of service systems involve visits by consumers who are "processed" at the headquarter points: restaurants, theaters, and other entertainment places of all sorts, schools, churches, hospitals, courts, museums, reference libraries. (Lending libraries, on the other hand, perhaps have more in common with systems in which headquarter points serve as storage places for commodities available on a rental basis; these in turn are closely related to retail store systems).

Another important category of service systems are those in which trips occur from the headquarter points out into the "field" in response to messages: fire crews, police, and rescue squads answering calls for aid, taxicab services, home repair services.

In the systems considered so far, determination of the quantity of services taken typically rests on the individual consumer, subject to pricing or rationing decisions by the managers of the headquarter points. However, in several systems of a control or housekeeping nature, the quantity of services supplied depends also on the headquarter manager: street cleaning and repair, public garbage collection, and, especially, inspection activity of all sorts--police patrollong, welfare, fire, health, housing, pollution control, etc.

Radio and television broadcasting constitute a rather special category. There are no transport costs varying with distance in the ordinary sense of the term; instead, the quality of reception varies by distance. The telephone and postal systems stand apart in that they provide connections between two members of the population, rather than direct services between a headquarter point and a consumer. The adequacy of the service system model as a representation of the real-world system is rather doubtful in these cases. (Broker services--e.g., employment agencies, real-estate companies--might perhaps be classified with the telephone and postal services).

We have thus briefly surveyed a wide variety of systems which might possibly be represented by the service system model.

How adequate this representation would be in any particular case can only be decided by a detailed examination of the system in question. This task will not be undertaken here. We will, however, make some comments on the shortcomings of the service system model which are inherent in the model itself and will appear in any application of it.

These shortcomings may be divided into intra- and inter-systematic. Under inter-systematic, we note first that multiple-purpose trips cannot easily be fitted into the service system framework; the possibility of visiting several headquarter points from different systems, which are located near each other, allows one to spread transport costs. Secondly, a clustering of headquarter points channels transport flows, and thus affects transport costs by allowing economies of mass transportation to be achieved, on the one hand, and by increasing congestion, on the other. Thirdly, production costs at headquarter points depend in part on the location of headquarter points of other linked systems. To catch these influences would appear to require the simultaneous determination of several service systems, together with the commuting problem of Chapter 1.

Under intra-systematic, we note first that production costs at a headquarter point depend not only on its own scale of operations, but on the location and scale of the other headquarter points of the system. These "external economies" arise in part from the use of common supplies, from the possibility of pooling outputs to meet demand fluctuations, from the exchange of information, etc. Secondly, it may not be possible to express total transport costs as a simple summation over the products of ideal weights by ideal distances. Thirdly, the concept "level of services to the consumer" is not always clearcut. In the case of commodity flow one has at least a simple quantitative measure (though benefits may depend on the level of neighbors' consumption as well as one's own). But how to measure the level of police protection or fire-fighting services? These will depend on density of patrolling, speed of response in emergencies, and various qualitative factors; there are strong "neighborhood effects" from these services, and benefits presumably depend on density of population, character of neighborhood, and other environmental variables.*

* The service system model might be reformulated to make benefits and some costs depend on population density and area, rather than being a summation over individual consumers. Cf. W.S. Vickrey "General and Specific Financing of Urban Services" in Public Expenditure Decisions in the Urban Community (H.G. Schaller, ed.) (Washington, Resources For the Future, 1963), and W.R. Thompson A Preface to Urban Economics (Baltimore, Johns Hopkins University Press, 1964), p.274.

With these reservations in mind we shall now develop further the service system model. It is hard to do much with the very general definition of service system given above, and so we shall examine special cases. A full list of assumptions will be given in the next section. Here we wish to examine one of them from the point of view of interpretive realism: the assumption that the quantity of services taken by all consumers is the same. At first glance this uniformity assumption seems very crude: first, because tastes differ among consumers, and second, because, even if tastes were uniform, the unit cost of services would rise with increasing distances from headquarter points, and this should lead to a fall-off of service levels with distance (unless we make the rather extreme assumption of completely inelastic demand). No doubt the uniformity assumption is very poor for many serviee systems. We merely wish to point out that for many other service systems it is much more plausible.

In the first place, a number of service systems operate under constraints which force everyone to receive uniform service levels. Compulsory education laws, for example, force all children in certain age brackets to attend schools with (roughly) the same number of hours per day and days per week for each. If we take hours of schooling as the measure of service level, we have an enforced uniformity for this population. Governmental service systems are often required to provide uniform services. For example, a fire crew will answer all alarms within its territory. (It may be argued that the extra delay in getting to a distant fire constitutes a lower level of service; but this may also be represented as greater transport cost for the same level of service). Police activities, street maintenance, garbage collection, public inspections may be required to perform uniform services.#

Writers on public finance discuss a category of commodities called "collective goods" which by their very nature are consumed equally by all. Whether there are such goods or not, it should be noted that the products of service systems are not collective; equal consumption of these must be provided for by special institutional arrangements, and will not occur otherwise.

Secondly, many service systems practice a policy of "freight absorbsion". That is, they make their service available at the consumption site on terms independent of the site's distance from the serving headquarter point, An example would be a retail store making free deliveries. (However, if the consumer has to make a shopping trip, or if the extra delay in making deliveries to more distant points is important, then there still remains a rising cost of services with distance to the consumer, though attenuated). Water supply and sewer services, electricity and gas, postal and telephone services are other systems which typically absorb freight, at least within a broad "local zone". If the cost to the consumer really is independent of his location, then we need only assume uniform tastes to get our uniform consumption result, not the much stronger assumption of inelastic demand.

3.5. The Scale and Spacing of Headquarter Points

We now formally investigate an abstract and simplified service system model. As discussed above, we assume that everyone receives the same level of services. Our space is a Euclidean plane, and population is assumed to be distributed uniformly upon it.* These assumptions together imply that the volume of services

* Incidentally, the assumption of uniform service levels increases the plausibility of the assumption of uniform population distribution. Ordinarily, population will tend to crowd up around headquarter points to reduce transport costs, but if uniform services are provided this centripetal attraction disappears. However, our aim in this section is not to justify our assumptions but to examine their consequences. (Centripetal attraction will be

taken up in the context of Thünen systems, in the next chapter).

provided per unit area is uniform throughout the plane. It is further assumed that all points are equally suited to be headquarter points; that is, there is a single function giving total production cost for any level of production of services, the same for all potential headquarter sites. Finally, it is assumed that tastes are uniform, in the sense that there is a single function giving benefit for any level of services received, the same function for all people. The unknowns of the problem are (1) the discrete set of points at which headquarters are to be placed, and (2) the scale of production at each of these points.

The criterion by which these unknowns are to be determined is: maximize total benefits minus the sum of total transport plus production costs. This, however, is not quite satisfactory as it stands, since it will presumably be infinite for several desighs --the plane, after all, and the total uniform population upon it are both infinite in our idealization. Instead, we take average _percapita_ benefits minus _per capita_ transport costs minus _percapita_ production costs as our criterion to be maximized. These are obviously equivalent for any given finite population, so that it is natural to use the second, which remains bounded even for an infinite population. There is still a third formulation which will prove useful: maximize total benefits minus costs _per unit area_. Since population density is fixed and uniform, this is obviously equivalent to the second criterion.

Finally, we assume (1) production levels at all headquarter points are the same, and (2) headquarter points are arranged in a hexagonal lattice. These assumptions are made with a certain diffidence, because--with mild conditions on benefit and production cost functions--it is likely that they could have been deduced from our previous assumptions. If we make the first assumption, and are given the average number of headquarter points per unit area and the level of production, percapita benefits and per capita production costs are already determined. The only remaining term in our criterion function is transport costs, and the optimal arrangement of headquarter points is the one minimizing this. We have a problem here very similar to our old friend, the headquarter location problem of Section 3.3. (They are identical except for the additional constraint here that all service areas have the same area). Our conclusion in that section, it will be recalled, was that the hexagonal lattice arrangement was plausible but not demonstrated.

In any case, these assumptions effect an enormous simplification. The number of essential degrees of freedom in the unknowns is reduced from infinity to two: (1) the production level at a headquarter point, and (2) the size of the "mesh" or spacing of the hexagonal lattice of headquarter points. These two numbers determine the per capita level of services, the size of the service area, total transport costs in a service area, and all other aspects of interest.

Our major interest lies in obtaining theorems in comparative statics; that is, in determining how the unknowns of the system respond to changes in the parameters of the system. We shall be concerned with two major parameters: (1) unit transport costs, and (2) population density. (A typical question would be "if population were denser, would headquarter points be spaced closer together or farther apart?") In the first part of our exposition we will be dealing with a fragmentary problem, and here it will be convenient to introduce still a third parameter:

(3) level of service per unit area.

For convenience we give a list of the symbols used:

γ service density; i.e. level of service per unit area.

θ transport cost parameter; the cost of conveying a unit weight of service over a unit distance.

y the scale of operation or level of production of services at a headquarter point.

C(y) average cost of production.

R the in-radius of a hexagonal service area (the radius of the inscribed circle); half the distance between adjacent headquarter points (see Figure 8, a blowup of Figure 7).

A the area of a hexagonal service area.

δ population density.

x level of services per capita.

B(x) per capita benefits.

P population served by a single headquarter point.

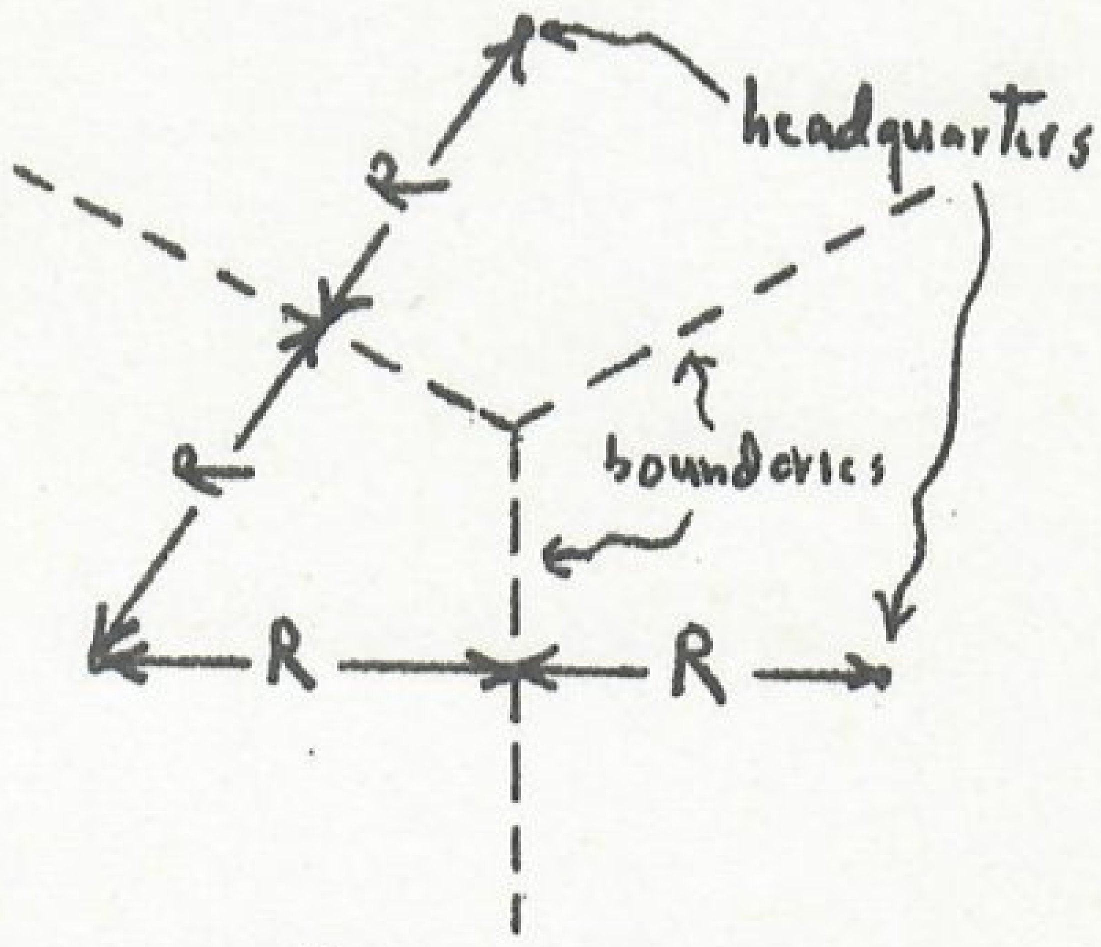

Figure 8

First we deal with the fragmented problem in which the service density, γ, is taken as given. We are to minimize the sum of production and transport costs per unit area by choosing R and y optimally. There is in fact just one degree of freedom here, given γ, since y and R have the relation

$$1) \qquad y = A\gamma = \frac{6}{\sqrt{3}} R^2 \gamma .$$

The whole system is a collection of congruent regular hexagonal service areas, each with a headquarter point at its center. Our criterion therefore leads us to minimize: total cost of production at a headquarter point plus total cost of transportation within a service area, all divided by the area of the service area. Total cost of production equals yC(y). To find the total cost of transportation, we note that an element of area dA within the service area and at distance z from the center contributes an amount $\gamma\theta z dA$ to total transport costs. Total transport costs is the integral of this over the hexagon. Taking the hexagon as the union of twelve 30-60-90 triangles, we evaluate the integral as

$$12\gamma\theta \int_0^{\frac{\pi}{6}} \left[\int_0^{R\sec\alpha} z^2\,dz\right] d\alpha \quad \text{(see Figure 9)}$$

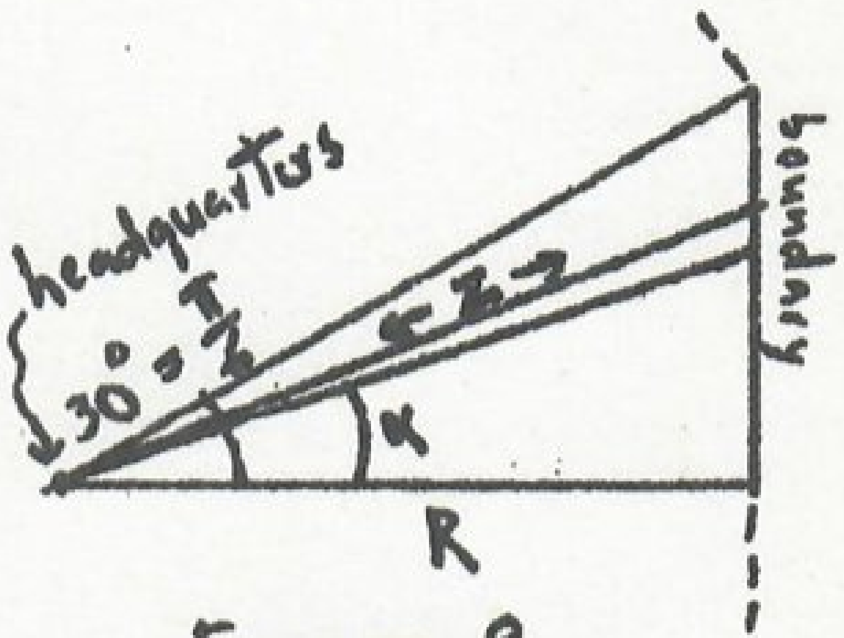

Figure 9

$$= \gamma\theta R^3 \left[2\tan\alpha\sec\alpha + 2\log(\tan\alpha + \sec\alpha)\right]_0^{\frac{\pi}{6}}$$

$$= \left(\frac{4}{3} + \log 3\right)\gamma\theta R^3.$$

We are, then, to minimize $\frac{1}{A}\left[yC(y) + \left(\frac{4}{3} + \log 3\right)\gamma\theta R^3\right]$.

When R and A are expressed in terms of y from relations (1), this becomes

2) $\underline{\gamma C(y) + a\theta\sqrt{y\gamma}} = M$, (where $\underline{a}$ stands for the constant $12^{-\frac{3}{4}}\left(\frac{4}{3} + \log 3\right)$).

Thus our minimand boils down to the rather simple form (2) with two terms, one involving average production costs, the other the square root of production.*

* Hexagonality of the service area is not at all crucial to the form of this result. Any packing of the plane with congruent service areas would give the same result, except for the constant.

If we knew C(y) we could minimize (2) explicitly over y, and thus express optimal production (and thus optimal spacing, from (1)) as a function of the parameters θ and γ. However, we can also discover properties of this function without specifying C(y). To do so, we make use of the following well-known result.

Lemma 1: Let $\bar{y}$ be an optimal solution to the problem:

minimize $h(\alpha, y)$, where

y is real, α is a real parameter, and $h(\alpha, y)$ is a real-valued function such that $\frac{\Delta^2 h}{\Delta\alpha\Delta y} > 0$ in the positive quadrant of the α, y plane; then $\frac{\Delta\bar{y}}{\Delta\alpha} \leqq 0$, if the α and $\bar{y}$ values are all positive.*

This lemma allows us to discover responses to parametric changes by observing the signs of cross-differences (or cross-derivatives, if $h(\alpha, y)$ possesses them), which are often obvious on inspection.

Let us now determine how scale and spacing respond to a change in unit transport costs, service density being held constant. We let the symbol $\left(\frac{\Delta\bar{y}}{\Delta\theta}\right)_\gamma$ stand for the change in $\bar{y}$ with respect to θ, with γ being fixed.

Theorem 1: $\left(\frac{\Delta\bar{y}}{\Delta\theta}\right)_\gamma \leqq 0$; $\left(\frac{\Delta\bar{R}}{\Delta\theta}\right)_\gamma \leqq 0$.

Proof: From the minimand (2), $\frac{\Delta^2 M}{\Delta\theta\Delta y} > 0$. Apply Lemma 1. This proves the first statement. The second follows from the fact that, given γ, R is a monotone increasing function of y, from (1). QED

Thus a fall in unit transport costs will increase the scale of operations at each headquarter point and expand service areas (or, in the limiting case, leave these unchanged).

* Cf. Samuelson Foundations of Economic Analysis, op. cit., p. 32, 49f, and Part I passim. A bar over a variable will always indicate the optimal value of that variable in this section.

Next, we hold unit transport costs fixed, and investigate how scale and spacing respond to changes in service density.

Theorem 2: $\left(\frac{\Delta \bar{y}}{\Delta \gamma}\right)_\theta \geqq 0$

Proof: Lemma 1 cannot be applied immediately. However, if we divide (2) through by γ, we get another criterion function which is equivalent to (2) (in the sense that the same optimal $\bar{y}$ results from both); for this criterion we get $\frac{\Delta^2(M/\gamma)}{\Delta\gamma\,\Delta y} < 0$. Now apply Lemma 1. QED

Thus a rise in service density leads to a rise in production levels. The effect on spacing, on the other hand, can go either way, depending on the average cost function $C(y)$. A condition on $C(y)$ can be given, which, however, has little intuitive appeal.

Theorem 3: If $\frac{d}{dy}\left(y\frac{dC}{dy}\right) > 0$ (<0) at $\bar{y}$, then $\left(\frac{\partial \bar{A}}{\partial \gamma}\right)_\theta \leqq 0$ $(\geqq 0)$.

Proof: It is convenient to let A be our unknown. Substituting $A\gamma$ for y in the criterion (2), and dividing through by γ, we get $C(A\gamma) + a\theta\sqrt{A} = M/\gamma$. This is to be minimized by $\bar{A}$. Taking cross-derivatives we get $\frac{\partial^2(M/\gamma)}{\partial\gamma\,\partial A} = C'(A\gamma) + A\gamma\, C''(A\gamma)$ $= \frac{dC}{dy} + y\frac{d^2C}{dy^2} = \frac{d}{dy}\left(y\frac{dC}{dy}\right)$. The theorem now follows from Lemma 1 in its local, differential, form. QED

Of the two cases in Theorem 3, the case $\frac{d}{dy}\left(y\frac{dC}{dy}\right) > 0$ seems more likely to occur in practice. For example, for the broad class of cost functions of the form $C(y) = \sum_i \alpha_i y^{\beta_i}$, with $\alpha_i > 0$ for all i, $\beta_i \neq 0$ for at least one i, we get $\frac{d}{dy}\left(y\frac{dC}{dy}\right) = \sum_i \alpha_i \beta_i^2 y^{\beta_i - 1} > 0$. We may therefore expect that, normally, a rise in service density will lead to a shrinkage of service areas.

We now turn our attention to the full cost-benefit problem. Service density, γ, is now an unknown, not a parameter. Taking its place is population density, δ. There are now two degrees of freedom for optimization. We shall find it most convenient to take as our two basic unknowns the scale of production, $\bar{y}$, as before, and the per capita level of services, $\bar{x}$. These, together with δ, then determine $\bar{\gamma}$, $\bar{R}$, $\bar{A}$, and $\bar{P}$. We are again most interested in the responses of these variables to chages in our two fundamental parameters, unit transport costs and population density.

It will be most convenient to use our second criterion, per capita benefits minus costs. The criterion (2) gives total costs per unit area, and division of this by δ converts it to total costs per capita $\frac{1}{\delta}\left[\gamma C(y) + a\theta\sqrt{y\gamma}\right]$. This is to be subtracted from per capita benefits, B(x), and the difference is now to be maximized. Upon substitution of $x\delta$ for γ, we get

3) $B(x) - x\,C(y) - \frac{a\theta}{\sqrt{\delta}}\sqrt{xy}$ as our maximand.

This is to be maximized over x and y. It will be noticed that, while we have two independent parameters, they enter (3) very conveniently in just one factor. We need only evaluate the effects on $\bar{x}$ and $\bar{y}$ of shifts in the single composite parameter $\frac{a\theta}{\sqrt{\delta}}$ to determine the effects of shifts in both θ and δ, by use of the chain rule. Let us abbreviate $\frac{a\theta}{\sqrt{\delta}}$ as β.

Lemma 1 is no longer serviceable, since we have two unknowns to be optimized jointly. Holding $\bar{y}$ constant, for example, may lead to a reversal of the response $\bar{x}$ makes to a shift in β, compared to the case where $\bar{y}$ is left free to adjust also to shifts in β. Instead, we make use of the following rather specialized result.

<u>Lemma 2</u>: Let $\bar{x}, \bar{y}$ be an optimal solution to the problem: maximize $M = B(x) - xC(y) - \beta\sqrt{xy}$. $\bar{x}$, $\bar{y}$, and β are all positive, and the ordinary first- and second-order conditions hold in a neighborhood of the maximizer. Then $\frac{d\bar{x}}{d\beta} \leqq 0$ and $\frac{d\bar{y}}{d\beta} \leqq 0$.

<u>Proof</u>: It will be convenient to proceed in three parallel steps, which are all combined at the end.

(1) Suppose we transform our variables, x and y, into ξ and η, where $\xi = xy$. β enters the revised maximand only in the term $-\beta\sqrt{\xi}$. By a result analogous to Lemma 1, it follows that $\frac{d\bar{\xi}}{d\beta} \leqq 0$ - i.e. $\bar{x}\frac{d\bar{y}}{d\beta} + \bar{y}\frac{d\bar{x}}{d\beta} \leqq 0$.

(2) By the second order conditions for a maximum,

$$\frac{\partial^2 M}{\partial x^2} = \frac{d^2 B}{dx^2} + \frac{\beta}{4}x^{-\frac{3}{2}}y^{\frac{1}{2}} \leqq 0 \text{ at } \bar{x}, \bar{y}.$$

$$\therefore \frac{d^2 B}{dx^2} < 0 \text{ at } \bar{x}.$$

$$\frac{\partial^2 M}{\partial y^2} = -x\frac{d^2 C}{dy^2} + \frac{\beta}{4}x^{\frac{1}{2}}y^{-\frac{3}{2}} \leqq 0 \text{ at } \bar{x}, \bar{y}.$$

$$\therefore \frac{d^2 C}{dy^2} > 0 \text{ at } \bar{y}.$$

(3) By the first-order conditions for a maximum,

$$\frac{\partial M}{\partial x} = \frac{dB}{dx} - C - \frac{\beta}{2}x^{-\frac{1}{2}}y^{\frac{1}{2}} = 0 \text{ at } \bar{x}, \bar{y}.$$

$$\frac{\partial M}{\partial y} = -x\frac{dC}{dy} - \frac{\beta}{2}x^{\frac{1}{2}}y^{-\frac{1}{2}} = 0 \text{ at } \bar{x}, \bar{y}$$

Multiply the second equation by $-\bar{y}/\bar{x}$, and add it to the first, to get

$$\frac{dB(\bar{x})}{dx} - C(\bar{y}) + \bar{y}\,\frac{dC(\bar{y})}{dy} = 0$$

Differentiate this equation with respect to β:

$$\frac{d^2B(\bar{x})}{dx^2}\,\frac{d\bar{x}}{d\beta} + \left[-\frac{dC(\bar{y})}{dy} + \frac{dC(\bar{y})}{dy} + \bar{y}\,\frac{d^2C(\bar{y})}{dy^2}\right]\frac{d\bar{y}}{d\beta} = 0,$$

or $$\frac{d^2B(\bar{x})}{dx^2}\,\frac{d\bar{x}}{d\beta} + \bar{y}\,\frac{d^2C(\bar{y})}{dy^2}\,\frac{d\bar{y}}{d\beta} = 0.$$

Now we put everything together. From step (2) the coefficient of $\frac{d\bar{x}}{d\beta}$ in this last equation is negative, and the coefficient of $\frac{d\bar{y}}{d\beta}$ is positive. Therefore both of these derivatives must have the same sign (or both must be zero). From step (1) it follows that this sign cannot be "+". $\therefore\ \frac{d\bar{x}}{d\beta} \leqq 0$ and $\frac{d\bar{y}}{d\beta} \leqq 0$.* QED

* Conceivably, this result could be generalized, or the proof made easier, by proceeding along the lines of Samuelson's "Generalized Le Chatelier Principle", according to which the responsiveness of a variable to a parametric shift rises if a constraint is removed. If $\bar{y}$ is held fixed one easily shows by Lemma 1 that $\frac{d\bar{x}}{d\beta} \leq 0$, and this should then still hold if the constraint on $\bar{y}$ is removed; similarly for the roles of $\bar{x}$ and $\bar{y}$ reversed. We have not investigated this. See Samuelson, op. cit., p.36-38.

Holding population density fixed, β varies in the same direction as unit transport costs. We then have

__Theorem 4:__ $\left(\frac{\partial \bar{x}}{\partial \theta}\right)_\delta \leq 0 \quad ; \quad \left(\frac{\partial \bar{y}}{\partial \theta}\right)_\delta \leq 0 .$

__Proof:__

$\left(\frac{\partial \bar{x}}{\partial \theta}\right)_\delta = \frac{\partial \bar{x}}{\partial \beta}\left(\frac{\partial \beta}{\partial \theta}\right)_\delta \leq 0,$ from Lemma 2 and the definition of β; similarly for the second statement. QED

Thus a fall in unit transport costs leads to a rise in the scale of production at each headquarter point and a rise in the level of services received by each person. Theorem 4 should be compared with Theorem 1; unit transport costs is the independent variable in both cases, but per capita service levels are held constant in Theorem 1 and allowed to vary in Theorem 4.

Holding unit transport costs fixed, β varies in the opposite direction to population density. This implies

__Theorem 5:__ $\left(\frac{\partial \bar{x}}{\partial \delta}\right)_\theta \geq 0 \quad ; \quad \left(\frac{\partial \bar{y}}{\partial \delta}\right)_\theta \geq 0 .$

__Proof:__

$\left(\frac{\partial \bar{x}}{\partial \delta}\right)_\theta = \frac{\partial \bar{x}}{\partial \beta}\left(\frac{\partial \beta}{\partial \delta}\right)_\theta \geq 0;$ similarly for $\bar{y}$. QED

Thus, for example, a rise in population density leads to a rise in the level of services received percapita (or in the limit leaves the level unchanged). This is perhaps the most interesting and least obvious of the four statements in Theorems 4 and 5. It may also be shown that the __net__ benefits (i.e. benefits minus costs) per capita rise with population density and fall with higher unit transport costs:

__Theorem 6:__ $\left(\frac{\partial M}{\partial \delta}\right)_\theta \geq 0 \quad ; \quad \left(\frac{\partial M}{\partial \theta}\right)_\delta \leq 0 ,$ where

$M = B(\bar{x}) - \bar{x}\,C(\bar{y}) - \beta\sqrt{\bar{x}\bar{y}}$

__Proof:__ $\left(\frac{\partial M}{\partial \delta}\right)_\theta = \frac{dM}{d\beta}\left(\frac{\partial \beta}{\partial \delta}\right)_\theta ,$ where $\frac{dM}{d\beta}$ allows $\bar{x}$ and $\bar{y}$

To adjust to variations in β. But under optimal adjustment of $\bar{x}$ and $\bar{y}$, $\frac{dM}{d\beta} = \frac{\partial M}{\partial \beta}^{*} = -\sqrt{\bar{x}\bar{y}} < 0 \quad \therefore \left(\frac{\partial M}{\partial \delta}\right)_{\theta} \geqq 0$; similarly for the second statement. QED

* Samuelson, loc. cit., p.34.

The intuitive explanation of the rise in net per capita benefits with increased population density is that a higher demand allows the headquarter point to take advantage of scale economies in production; (average cost must be falling at equilibrium).

We turn now briefly to examine the responses of other variables to shifts in our parameters. Unfortunately--with one exception--these cannot be determined from the results of Theorems 4 and 5 alone. The exception concerns the response of service density to a change in unit transport costs, and for this we get

$$\left(\frac{\partial \bar{y}}{\partial \theta}\right)_{\delta} = \left(\frac{\partial (\bar{x}\delta)}{\partial \theta}\right)_{\delta} = \delta \left(\frac{\partial \bar{x}}{\partial \theta}\right)_{\delta} \leqq 0, \text{ from Theorem 4.}$$

Let us take, for example, $\bar{P}$, the population of a service area. $\bar{P}$ equals $\bar{y}/\bar{x}$, and since $\bar{x}$ and $\bar{y}$ respond in the same direction to parametric shifts, the effect on $\bar{P}$ cannot be determined from this qualitative knowledge alone. We do, however, have the following result.

<u>Theorem 7</u>: Under the assumptions of Lemma 2,

$$\bar{x}\,\frac{d^2B(\bar{x})}{dx^2} + \bar{y}^2\,\frac{d^2C(\bar{y})}{dy^2} \geqq 0 \ (\leqq 0) \text{ implies that}$$

$$\left(\frac{\partial \bar{P}}{\partial \theta}\right)_{\delta} \geqq 0 \ (\leqq 0), \text{ and } \left(\frac{\partial \bar{P}}{\partial \delta}\right)_{\theta} \leqq 0 \ (\geqq 0).$$

Proof: Given δ, θ determines β, β determines $\bar{x}$ and $\bar{y}$, and these determine $\bar{P}$, so $\left(\frac{\partial \bar{P}}{\partial \theta}\right)_\delta = \frac{d\bar{P}}{d\beta}\left(\frac{\partial \beta}{\partial \theta}\right)_\delta$. The second factor is positive, so we need only determine the sign of $\frac{d\bar{P}}{d\beta} = \frac{d(\bar{y}/\bar{x})}{d\beta} = \frac{1}{\bar{x}^2}\left(\bar{x}\frac{d\bar{y}}{d\beta} - \bar{y}\frac{d\bar{x}}{d\beta}\right)$.

In the course of proving Lemma 2, we found that $\frac{d^2B(\bar{x})}{dx^2}\cdot\frac{d\bar{x}}{d\beta} + \bar{y}\,\frac{d^2C(\bar{y})}{dy^2}\cdot\frac{d\bar{y}}{d\beta} = 0$. Solving for $\frac{d\bar{x}}{d\beta}$, and substituting, we get

$$\frac{d\bar{P}}{d\beta} = \frac{1}{\bar{x}^2}\left[\bar{x}\frac{d\bar{y}}{d\beta} - \bar{y}\left(-\frac{\bar{y}\,\frac{d^2C(\bar{y})}{dy^2}\,\frac{d\bar{y}}{d\beta}}{\frac{d^2B(\bar{x})}{dx^2}}\right)\right]$$

$$= \frac{\frac{d\bar{y}}{d\beta}}{\bar{x}^2\,\frac{d^2B(\bar{x})}{dx^2}}\left[\bar{x}\,\frac{d^2B(\bar{x})}{dx^2} + \bar{y}^2\,\frac{d^2C(\bar{y})}{dy^2}\right].$$

$\frac{d\bar{y}}{d\beta} \leqq 0$, and $\frac{d^2B(\bar{x})}{dx^2} < 0$, so the factor outside the brackets is $\geqq 0$, so the sign of $\frac{d\bar{P}}{d\beta}$ is the same as the sign of the bracketed expression, and the first statement follows. The second follows from $\left(\frac{\partial\beta}{\partial\delta}\right)_\theta < 0$.

QED

As a corollary, the direction of response of headquarter spacing to changes in unit transport costs can be conditionally ascertained, since $\left(\frac{\partial \bar{A}}{\partial \theta}\right)_\delta = \left(\frac{\partial(\bar{P}/\delta)}{\partial \theta}\right)_\delta = \frac{1}{\delta}\left(\frac{\partial \bar{P}}{\partial \theta}\right)_\delta$, so that service areas expand if and only if population served by a headquarters expands. These results offer a remarkable contrast to Theorem 1. There, a fall in unit transport costs must result in expanded service areas (or no change, in the limit); here, there is the possibility that a fall in transport costs actually leads to a shrinkage of service areas. The mechanism is that, while scale of production expands, level of services received per person expand in a greater ratio, so that the population and area served by a headquarter point falls. The possibility seems counter-intuitive, and might be labeled the "spatial Giffen paradox."

So far we have not actually demonstrated that this possibility can be realized, and conceivably the crucial expression $\bar{x}\frac{d^2B(\bar{x})}{dx^2}+\bar{y}^2\frac{d^2C(\bar{y})}{dy^2}$ could never be made positive by any choice of benefit and cost functions. The simplest way of refuting this contention is by counter-example.

Let $B(x) = m\sqrt{x}$, and let $C(y) = n/y$, where m and n are any positive constants. One may verify that the optimal solutions are $\bar{y} = \frac{m^2}{4\beta^2}$, $\bar{x} = \frac{m^6}{256\,n^2\beta^4}$, $\bar{p} = \bar{y}/\bar{x} = \frac{64\,n^2\beta^2}{m^4}$.

$\bar{P}$ obviously varies directly with β, so the "spatial Giffen paradox" obtains. (The functions B(x) and C(y) were chosen for simplicity. The particular choice for C(y) makes marginal production costs zero; this is not at all crucial. One has only to choose a sufficiently "flat" benefit function and a sufficiently "curvy" cost function to generate the paradox.)

We will end the investigation of this interesting and important problem at this point, though the questions that may be posed are far from being exhausted. First, the effects of shifts in population density have not all been worked out. Second, there are several other fragmented problems that might be investigated; for example, the pattern of locations may be given in advance, so that A and R are no longer unknowns; or, the scale of operations at headquarter points may be given, so that y becomes a parameter. (Both these cases represent institutional constraints which are by no means far-fetched). Finally, one may expand the problem by introducing new parameters--for example, shift parameters in the benefit and cost functions, like "m" and "n" in the counterexample just given.

3.6. Löschian Equilibrium

This section differs from the remainder of Chapter 3 in that it refers to the interaction of several decision-makers, rather than the choice of just one. The model examined here is identical to that of the last section with the following exceptions.

(1) The level of services received by a person need no longer be uniform. Instead, it depends on a demand function giving the level of services for any local price (that is, the price of the service at the place where the consumer is located). It is assumed that all consumers have the same demand function.

(2) The local price system is assumed to obey the efficiency postulates of Section 2.4, so that, in a demand system, the local price at any point in an effective demand area equals the headquarter price plus transport costs. Headquarter prices are set to clear the market at the chosen production levels.

(3) Each headquarter point is thought of as an individual firm, which sets its scale of operations to maximize its profits, profits being production times headquarter price, minus total

production costs. Each firm is also free to re-locate its headquarter point.

(4) Free entry and exit of firms is permitted. Formally, the effect of this assumption is that no firm makes negative profits, and that there is no unused site at which a firm could locate and make positive profits at some production level. Lösch goes further, and assumes that the average size of service area is minimized, subject to the constraint of non-negative profits for all firms. (Since we are on an endless plain, one must interpolate some kind of limit process to give meaning to this statement, as discussed in Section 3.2.)

The remaining assumptions: uniform population distribution, uniform cost conditions, Euclidean metric, are as in the previous section. This ensemble is a (simplified) version of the Löschian equilibrium system.*

* A. Lösch The Economics of Location, op. cit., p.94-97.

Next, we assume with Lösch that firms will arrange themselves in a hexagonal lattice. Possibly this is deducible from the assumption that the average size of service area is minimized, plus some unknown conditions on demand and cost functions. We have already noted that the usual arguments are not conclusive in the case of the headquarter location problem, and the present problem is far more complicated. In any case we may merely take this as an extra assumption. It is also assumed that prices and scale of production at all headquarter points are equal, and that equilibrium profits are zero for all firms. (Again, some of these statements may be deducible from previous assumptions.

The logical structure of the Löschian system has never been thoroughly investigated).*

The service areas are, of course, congruent regular hexagons. We now recall the distinction made, in Section 2.4, between effective and potential market areas. The potential market area of a headquarter point is the set of all sites for which that point is the most economical source (for potential demand areas) or sink (for potential supply areas). These potential market areas partition the plane completely, and are known as service areas in the special case where prices at all headquarter points are equal (cf. Secćion 3.3). The effective market area of a headquarter point, on the other hand, consists of those sites which are actually linked to the headquarters by flows; this is a subset of the potential market area--a proper subset if some sites in the potential market area are not linked to the headquarter point.

The question now arises: what are the shapes of the effective market areas in a Löschian system? It seems to have been implicitly assumed that these coincided with the potential market areas until Mills and Lav showed that this need not be the case.* Their basic argument is ingenious and quite simple.

* For critiques of the Löschian system see M.J. Beckmann "Some Reflections onLösch's Theory of Location" Regional Science Association, Papers and Proceedings 1:N1-N8, 1955 and "The Economics of Location" Kyklos 8:416-421, 1955; Isard, op. cit., p.48f,270ff.; B.J.L. Berry and W.L. Garrison "Recent Developments of Central Plase Theory" Regional Science Association, Papers and Proceedings 4:107-120, 1958; E. von Böventer "Towards a Unified Theory of Spatial Economic Structure" Regional Science Association Papers 10:163-187, 1963.

* E.S. Mills and M.R. Lav "A Model of Market Areas with Free Entry" Journal of Political Economy 72:278-288, June 1964, p.283f.

Assume that the demand function is non-increasing and continuous, and that there is some cut-off price beyond which demand goes to zero. Take a single firm with no nearby competitors. It is easy to see that its effective demand area must be a circular disc. (If the headquarter price is set above the cut-off price, the disc shrinks to nothing. In a general metric, the effective demand area would be an out-sphere, in the terminology of Section 2.5). By varying its headquarter price down or up, the firm will make its demand area expand or contract, respectively. Assume there is some unique price at which total profit is maximized, and that this profit is positive. This price, and the corresponding effective demand area, are the equilibrium conditions for our isolated firm. Let us call this the full area case.

Now let us suppose that part of the circular disc becomes unavailable to our firm--perhaps because of the encroachment of competing firms. Call this the truncated area cade. Suppose the firm adjusts optimally to this contretemps so far as pricing is concerned (but maintains its old location). It is fairly obvious that the profit achieved by the firm in the truncated area case fallsbelow the profit achieved in the full area case. (Proof: let p_t and y_t be optimal price and production in the truncated case; let $\tilde{p}$ be the price at which y_t will be sold in the full case; there always is such a $\tilde{p}$ and, furthermore, $\tilde{p} > p_t$; thus revenue rises while costs remain the same; the optimal price in the full area case, p_f, cannot do worse than $\tilde{p}$, and therefore the full area profit exceeds the truncated area profit. QED)

Suppose next that the production cost function in these two cases contains a set-up cost parameter. An increase in set-up costs by an amount k increases total production costs by k for all levels of output (except output zero, for which cost is always zero). Such a change has no effect on optimal policy, provided profit remains positive, but merely reduces total profit by an amount k.

It follows from the preceding theorem that one can always choose a set-up cost so that profit in the full area case remains positive while profit in a certain truncated area case becomes negative. This observation is the key to showing that the effective demand areas need not cover the plane. For covering to occur, the circular discs would have to be compressed into regular hexagons, by the encroachment of the six firms adjacent to a given firm in the honeycomb lattice. By suitable choice of set-up cost, one could make profit in this case negative, for all possible combinations of pricing and spacing that fill the plane completely with effective demand areas. Take the most profitable hexagon size, and increase the lattice spacing enough so that the circumscribed circles about these hexagons can fit in the plane without overlapping. If we have chosen our set-up cost properly, the profit for this case will be positive. Thus we have shown that arrangements giving positive profits may exist even though all arrangements filling the plane completely with effective demand areas make negative profits, hence are not viable.

drawing from the industry. Any encroachment whatsoever of firms on each other's demand areas will make profits negative and drive them from the industry. The effective demand areas must therefore remain circular discs of radius R. The result is clearly Figure 10: the firms can not get closer to each other, and need not remain farther apart.

Full space-filling hexagonal effective demand areas also occur--for example, in the case of completely inelastic demand functions.

Given our assumptions of hexagonal lattice arrangement and equal prices at all headquarter points, the remaining possibilities are easily found. They are all of the form of Figure 11, where the effective demand area of a typical headquarter point O is depicted. It is bounded by twelve pieces. A,B,C,D,E, and F are equal line segments, perpendicular to, bisecting, and bisected by, the spokes connecting O to its six adjacent headquarter points; a,b,c,d,e, and f are equal circular arcs, all with center at O. The arcs are part of the rim of the circle along which local price hits the cut-off point. This rim is interrupted by the encroachment of adjacent firms, and here the effective demand area follows the service area boundaries. The shaded areas, bounded by curvilinear triangles, are the portions of the plane uncovered by effective market areas.

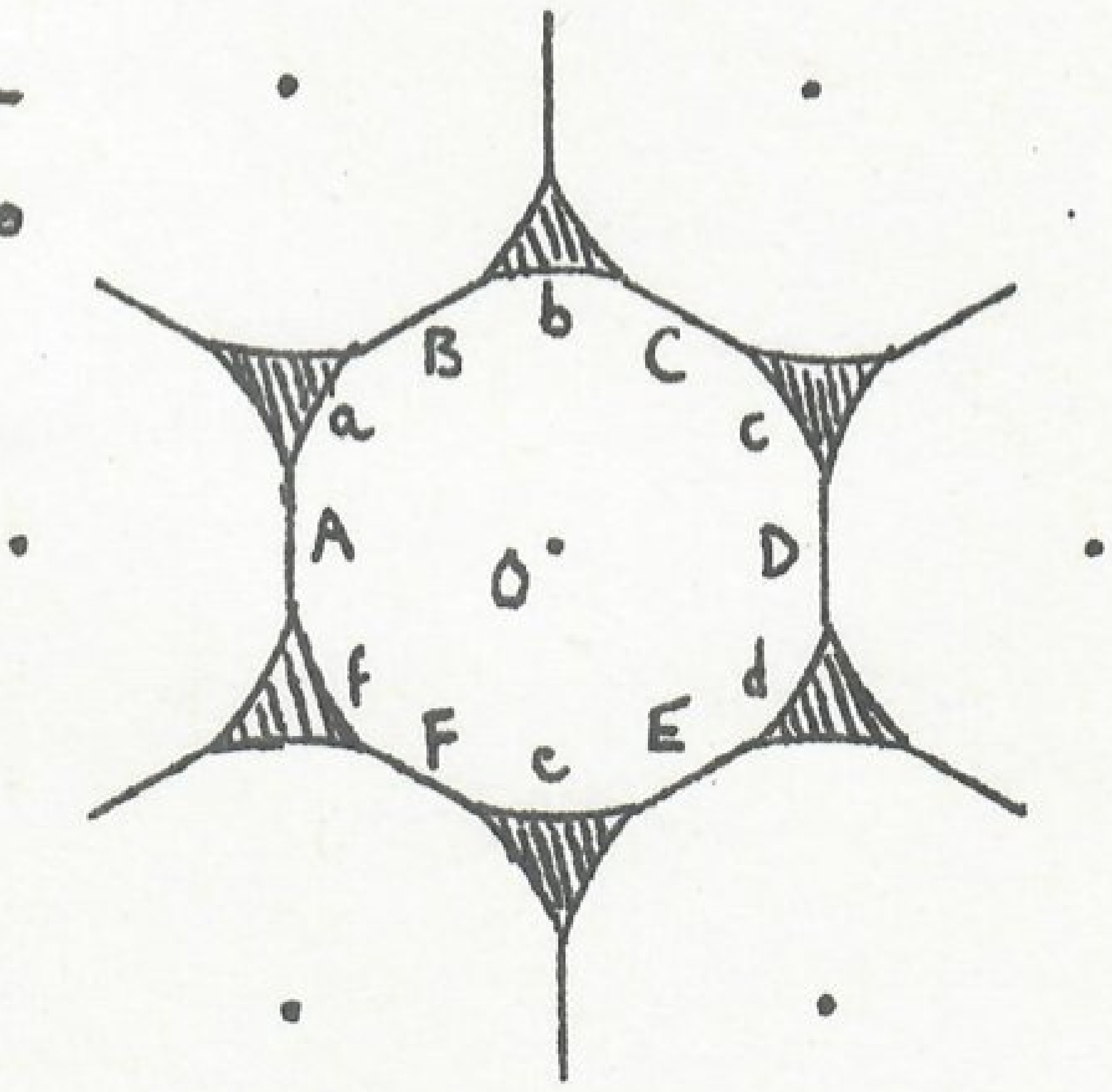

Figure 11

Given the lattice, Figure 11 retains one degree of freedom: the angle subtended by an arc. As the angle approaches zero, the arcs disappear, the shaded areas disappear, the line segments join to form a regular hexagon, and we are in the case of space-filling hexagonal effective demand areas. As the angle approaches 60°, the line segments disappear, the arcs join to form a full circle, and we are in the circular network of Figure 10. Thus these special cases are just the termini of a one-dimensional continuum of possibilities.#

The basic geometrical fact which underlies the possibility of areas of the plane going unserved is that the full areas--which are circular discs--cannot be arranged to cover the plane without overlapping. This is a property of the two-dimensional Euclidean metric, and with other metrics it need not occur. For example, in one-dimensional Euclidean space (i.e. the real line) the "spheres"are simply intervals, and these can be fitted snugly so that no finite interval is left uncovered. Even in two dimensions this may occur. Let us consider, for instance,

Mills and Lav, p. 285, have a diagram very similar to Figure 11. In searching for other possible shapes besides hexagons and circles, they try out regular dodecagons (and suggest regular 6n-gons as other possibilities). Here their penchant for polygons seems to have led them astray. Every other side of the dodecagon --a,b,c,d,e, and f--adjoins unserved territory, and therefore must bulge out into a circular arc. We would conjecture that Figure 11, and its two limiting cases, are the only effective demand area types in the Löschian system.

the important rectangular, or city-block, metric. This is defined by $d(L',L'') = |X'-X''| + |Y'-Y''|$, where X',Y' are the ordinary cartesian co-ordinates of the point L', and X",Y" are the ordinary cartesian co-ordinates of the point L", for a fixed pair of axes.*

For the rectangular metric it is easy to show that the "spheres" are ordinary squares, tilted 45° so that their diagonals are parallel to the axes. A system of service areas having this shape can cover the plane exhaustively (see Figure 12). The lattice of headquarter points corresponding to this system is built up from isoceles right triangles (one of these is shaded in Figure 12), rather than from equilateral triangles. We end our discussion with two conjectures. (1) The optimal arrangement of headquarter points, for the headquarter location or Löschian problems, is the checkerboard lattice of Figure 12, if the metric is rectangular. (2) In Löschian equilibrium the effective demand areas always cover the plane, if the metric is rectangular.

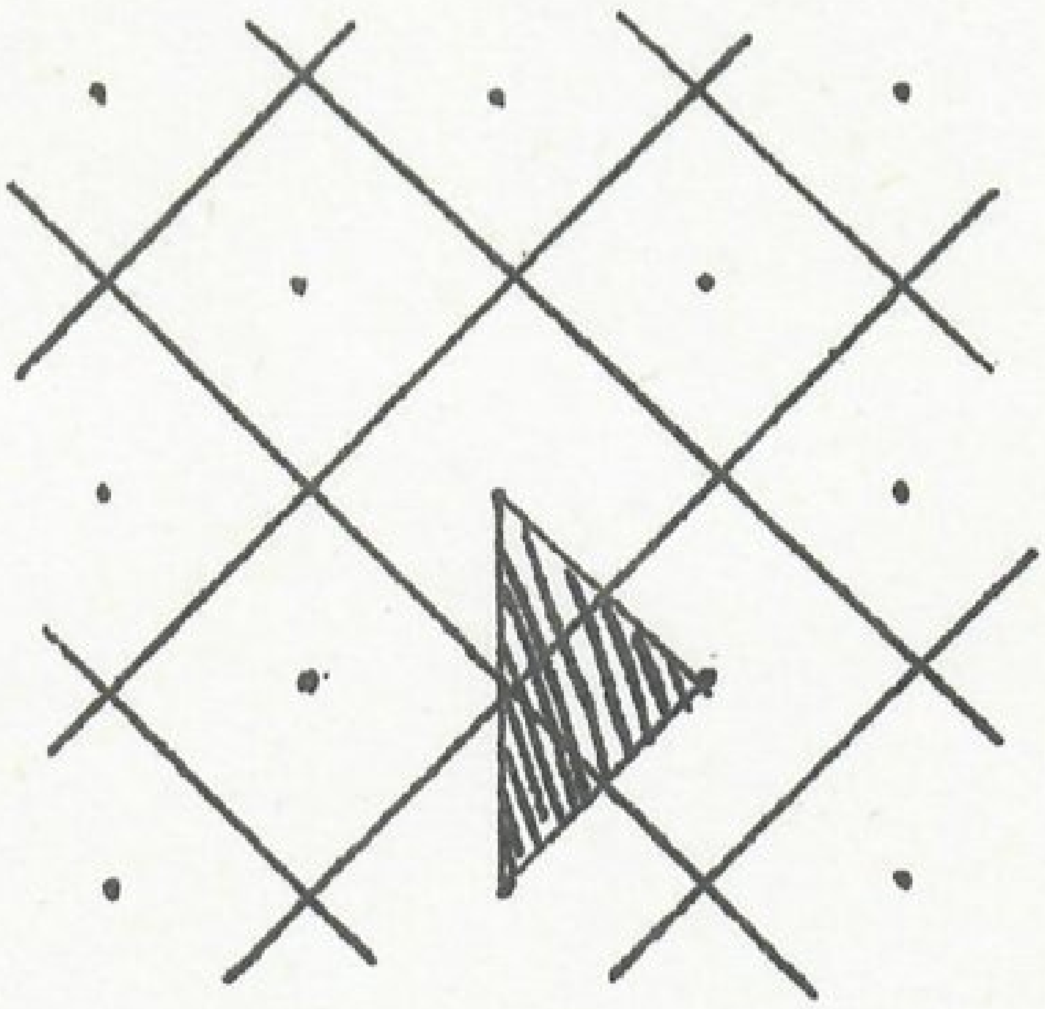

Figure 12

* The names arise from the fact that this metric can be generated from the Euclidean metric by allowing movements only parallel to the axes, as would o^cur in a city made up of rectangular blocks (of infinitesimal size), or as a rook moves in chess.

4. The Theory of Thünen Systems

4.1. Thünen Systems

Thünen's original work appeared in 1826.* The central model was designed to explain the location of agricultural land uses. It postulated an economy isolated from the outside world, consisting of a single city trading with its agricultural hinterland. Distance from the city was the major determinant of land use, profitability of the various competing uses being influenced by transport costs to and from market.

We shall be interested in a much wider class of models, all of which, however, share certain essential features with the original Thünen model. Suppose we have a metric space which is symmetric (i.e. $d(L,M)=d(M,L)$). In this space there is one distinguished point, called the nucleus. Land use at any site depends only on the distance between that site and the nucleus. Any pattern satisfying these conditions will be called a Thünen system.

In the original Thünen model, it is of course the city, at the center of the isolated state, which plays the role of the nucleus. It need not be assumed that a Thünen system is cut off from the rest of the world. Nor need the land uses be only agricultural in character.

We now give some concrete illustrations of situations which might be satisfactorily represented as Thünen systems. The best one can hope for, of course, is that certain broad features of the real-world situation are well-approximated by the model. It

* J.H. von Thünan Der Isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie (Hamburg, 1826).

would be absurd, for example, to try to locate the exact geometric point on the surface of the Earth which corresponds to the nucleus of the system. The nucleus may be a house, or a neighborhood, or a central business district, or an entire city (the original interpretation), or even wider regions: metropolitan areas, industrial belts, entire nations. The important point is that--for the particular situation in hand--these entities or regions are so small a part of the whole that they may be thought of as points.* It would also be futile to expect an exact ring pattern of land uses surrounding the nucleus to manifest itself in a real-world situation.

With these points in mind, one finds a rich variety of situations which approximately fit the Thünen pattern of concentric zones of land uses. In addition to city-agricultural hinterland systems, one has city-suburb systems. The internal structure of cities may sometimes be represented by a CBD surrounded by concentric zones of different classes of residential housing. Continuing down in scale, we may find the fields surrounding a small village differentiated in use by distance from the village. An individual farm may be organised as a miniature Thünen system, the nucleus being a cluster of farmstead and barns. Public assemblies, as for speeches, plays or boxing matches, may be thought of as Thünen systems in which the "land uses" are different qualities of seating, and the center of attention is the

* Think of a city a few miles in diameter surrounded by an agricultural hinterland stretching scores of miles. Cf. the discussion of Weberian activities, Section 3.2.

nucleus. At the other end of the scale, a nation may have a dominant political or economic center which serves as nucleus for organizing the whole economy into a big Thünen system. In the late 19th - early 20th Century, Western Europe may have become to some extent such an organizing center for the entire world.*

* On city-hinterland relations see E. deS. Brunner and J.H. Kolb Rural Social Trends (New York, McGraw-Hill, 1933), Chap. 5; D.J. Bogue The Structure of the Metropolitan Community (Ann Arbor, 1949); A.H. Hawley The Changing Shape of Metropolitan America (Glencoe, Free Press, 1956; and a series of papers by K.A. Fox on "Functional Economic Areas". The "concentric circle" theory of internal city structure is the product of the "Chicago school" of urban ecology; see E.W. Burgess "Urban Areas" pp.113-138 of Chicago: An Experiment in Social Science Research (T.V. Smith and L.D. White, eds.; Chicago, University of Chicago Press, 1929). The relevance of the Thünen model to urban land use was pointed out by W. Isard Location and Space-Economy, op. cit., Chap. 8, Appendix. Possible applicability of the Thünen model to individual villages and farms was mentioned by A. Lösch The Economics of Location, op. cit., p.62note45. For very large scale systems, see J.Q. Stewart "Empirical Mathematical Rules Concerning the Distribution and Equilibrium of Population" Geographical Review 37:461-485 July, 1947, for description, and, for some theoretical speculations, A. Weber "Die Standortslehre und Die Handelspolitik" translated by W.F. Stolper, International Economic Papers, #8. The quantity of literature that might be cited in addition is truly enormous.

In judging the "goodness of fit" of a Thünen system to a situation, one must remember that, even if the fit were perfect, the concentric zones need not look very circular in the geographical sense. The system is defined in terms of ideal distances which need bear no close relation to geographical distances. Formally, the situation is as follows. Take any "sphere" having the nucleus as a center. (Since the distance function is symmetric, in-spheres coincide with out-spheres of the same radius). The "rim" of this sphere consists of all points which are at a certain fixed distance from the nucleus, that distance being the radius of the sphere. If we do indeed have a Thünen system here, all these points will have been devoted to the same land use. Conversely, if for every sphere about a certain point, the rim of that sphere is devoted to one homogeneous land use, we have a Thünen system with the common centerpoint as nucleus. Now the rims of spheres may be quite irregular objects in the geographical sense (e.g. elongated along transport arteries, oreven split into several pieces, as in Figures 4 and 5, Section 2.5). The zones will follow these irregularities, but the whole will remain a Thünen system in good standing for all that.

We shall be analyzing two special kinds of Thünen systems. These are defined by the structure of resource flows within the system. The first will be called an <u>entrepôt system</u>. This one is characterized by the fact that the only flows which occur in the system are between a site and the nucleus (in either direction). Any output from a site is transported to the nucleus, and any input to a site is transported from the nucleus. There are no <u>direct</u> flows between two non-nuclear sites, L and M, but one may have a resource flow from L to the nucleus N, followed by a flow of the same resource frmm N to M. If we assume that flows follow routes which are geodesics, this means that the nucleus is *between* every pair of different sites, such as L and M, for which this phenomenon occurs. In fact, the simplest way of characterizing entrepôt systems--and the source for the name--is the condition that the nucleus is between every pair of different sites, i.e., that it is an *entrepôt* for the system, in the language of Section 2.1. Furthermore, all "foreign trade" is to be channeled through the nucleus; that is to say, there is no direct contact between non-nyclear sites and the outside world. This follows from the condition that the nucleus is a *gateway* between the system and the outside world, again in the language of Section 2.1. These two conditions--the nucleus as entrepôt for domestic affairs and as gateway for foreign affairs--serve as an alternative definition for an entrepôt system.

The second sub-case will be called a *direct-linkage system*. Again we assume that flows follow geodesic routes, as follows from the efficiency postulates of Section 2.4. Direct-linkage systems are characterized by the following property of the metric. For every point L at distance d from the nucleus, and for every d' such that $0 < d' < d$, there is a point M at distance d' such that (L,M,N) is a geodesic, N being the nucleus. (N,M,L) is of course also a geodesic, since the metric is symmetric. That is to say, the geodesic between any site and the nucleus passes through the rim of every sphere of radius smaller than d(L,N). (By contrast, in the entrepôt case the geodesic passes through no intermediate points; it is the direct route between L and N, for any point L.) In general, this means that the entrepôt assumption no longer holds; the gateway assumption is maintained, however: any contact with the outside world is channeled through the nucleus exclusively.

A simple example will illustrate these two cases. Suppose the land-use at distance d produces a certain commodity as an output, while the land-use at distance d' uses the same commodity as an input. In the entrepôt case there would be a flow of the commodity from the rim at distance d inward to the nucleus, and a counter-flow of the same commodity outward from the nucleus to the rim at distance d'. This could not occur under direct linkage, for it would involve cross-hauling, which violates the efficiency postulates. Through any rim a commodity can be flowing inward to the nucleus, or outward from the nucleus, or neither, but not both at once. The nucleus is by-passed (except for foreign trade) and producers and consumers in the field may be directly linked to each other--hence the name.

Under what conditions may we expect the formation of Thünen systems, or approximations thereto? When will the entrepôt case occur? Direct linkage? Neither? The answers to these questions will help put our definitions and interpretations in perspective.

The most frequent situation in which a Thünen system becomes established seems to be when an activity of great attractive power locates at a site not too close to another of comparable power. The competition for nearby land to reduce transportation costs to or from the attractive site then leads to the formation of rings or zones surrounding that site, and it becomes the nucleus of the resulting Thünen system. In this case the system is simply the market area of a headquarter point, in a way. (It should be noted, though, that the analysis of Sections 3.5 and 3.6 depends on the assumption of a single commodity flow. In

Thünen systems one deals, in general, with heterogeneous, multi-commodity flows). But not all market areas are Thünen systems; the headquarter activity may be too inconsequntial to have an effect on its environment. In this case the service system superimposes itself on the established pattern without disturbing it. The focus of attention is, in any case, completely different than it was in Chapter 3. Here the emphasis is on the determination of the pattern in the "field"; what goes on at the headquarter point or nucleus is taken as exogenous. In the last chapter precisely the opposite was the case.

A second situation is where a convenient point of access to the transportation system assumes the role of nucleus and a Thünen system develops around it. Examples are, a natural harbor, the head of navigation on a river, a railway station. Subsequent development will generally lead to an intense concentration of activity in the immediate vicinity of the nucleus, and this situation then merges into the situation discussed above.

An off-shoot of the second situation occurs when an original point of entry into a territory becomes the nucleus of a Thünen system; for example, a trading post or fortification. Here the system develops from the inside out, so to speak, waves of settlers pushing outward from the center. In the other cases, the system may develop instead by a "condensation" of the surrounding pre-existent population, as in rural-urban migration.

Once the original pattern develops, it tends to be self-perpetuating for several reasons. Transportation construction tends to concentrate along the routes of most intense traffic flow. These routes will be between the nucleus and the various hinterland points, so construction takes the form of a number of spokes radiating from the nucleus in several directions. This re-inforces the attractive power of the nucleus and tends to "freeze" the pattern (but also alters the metric, and therefore leads to some re-shaping of the Thünen zones). This is the situation in which "all roads lead to Rome". The growth of population, the rise of a substantial capital plant around the nucleus, the development of ties of trade and employment between nucleus and hinterland, the growth of sentimental attachments to the land--all conspire in the same direction.

We turn now to our two special cases, the entrepôt and direct-linkage systems. Once again let us emphasize the idealized and approximative character of these representations. To say that a certain metropolitan area is well represented as an entrepôt system is not to say that there is a certain lot upon which the entire traffic of the community converges. It <u>is</u> to say that the great bulk of traffic in the community ás between nucleus and hinterland, where the nucleus is spatially very compact and a small fraction of the entire land area.

By and large, the entrepôt model characterizes "young" Thünen systems, while the direct-linkage model characterizes "mature" Thünen systems. In the beginning, Thünen rings are unlikely to form without a strong centripetal pull of some sort, and a consequent polarization of traffic into the entrepôt pattern. The building of a radial transportation system re-inforces the entrepôt-like characteristics of the nucleus, and, in fact, if radial arteries were the only available means of transportation, and if they were all through-roads, with no intermediate turn-offs, the metrical requirements for an entrepôt system would be literally fulfilled: the nucleus would be between every pair of different sites. (See Figure 1). This condition is usually not even approximately fulfilled, but there is another factor at work which tends to make the system behave as if the entrepôt metric were embodied in the transportation system. This is the "middleman" function played by the nucleus. Suppose there are several potential buyers and sellers of a commodity scattered about the countryside and unaware of each other's locations. A market at the nucleus, with which all are acquainted, serves to bridge this information gap. From the point of view of an omniscient observer, buyers and sellers all going to a central market involves cross-hauling and excess transportation costs. From the point of view of the participants themselves, the shortest route to their trading partners goes through the market. In brief, ignorance leads to an entrepôt metric. (Economies of bulk transactions might also produce this centralization even if information were perfect).

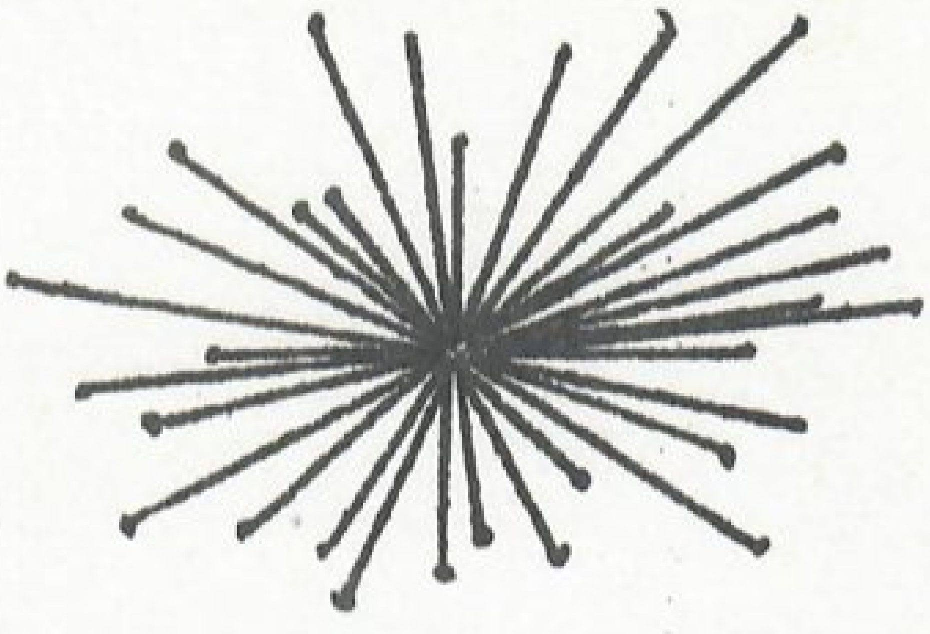

Figure 1

As the system grows, a number of forces arise to disrupt the entrepôt pattern. (1) The farther out potential buyers and sellers live, the more circuitous becomes trading at the center, and the more is to be gained by trading in the field and by-passing the nucleus. (2) Potential trading partners learn of each other's locations in time, which tends to diminish the role of the broker at the center. (3) Increased volumes of traffic and trade permit new headquarter points of all sorts to spring up in the field, having a sufficient demand to become viable. (4) Increased traffic creates ever more congestion at the nucleus. (5) The heavy investment in capital plant around the nucleus, which initially is an attractive force, may become a net dispersive force on balance, as structures become aged or obsolescent, and as it becomes more difficult to find uncluttered parcels of sufficient size for new enterprises. (6) Contact with the outside world, perhaps originally channeled completely through the nucleus, because of limited information about and access to the hinterland, tends to by-pass the nucleus as time goes on, information improves, and the transportation network begins to connect the hinterland directly to the outside world.

The direct-linkage model is perhaps a good representation for the "mature" condition resulting from the action of the forces listed above. The original entrepôt pattern becomes eroded, although the Thünen ring structure persists for some time by historical inertia. Perhaps there is a further "senile" stage in which even the ring structure becomes disrupted. In the mature stage, as here conceived, foreign trade is still largely a nuclear prerogative. Domestic trade is handled at numerous centers in the field in addition to the original nucleus. The flow of traffic is still largely radial in pattern (i.e. it moves between distance zones, relative to the nucleus), though much of it never reaches or stems from the nucleus itself.

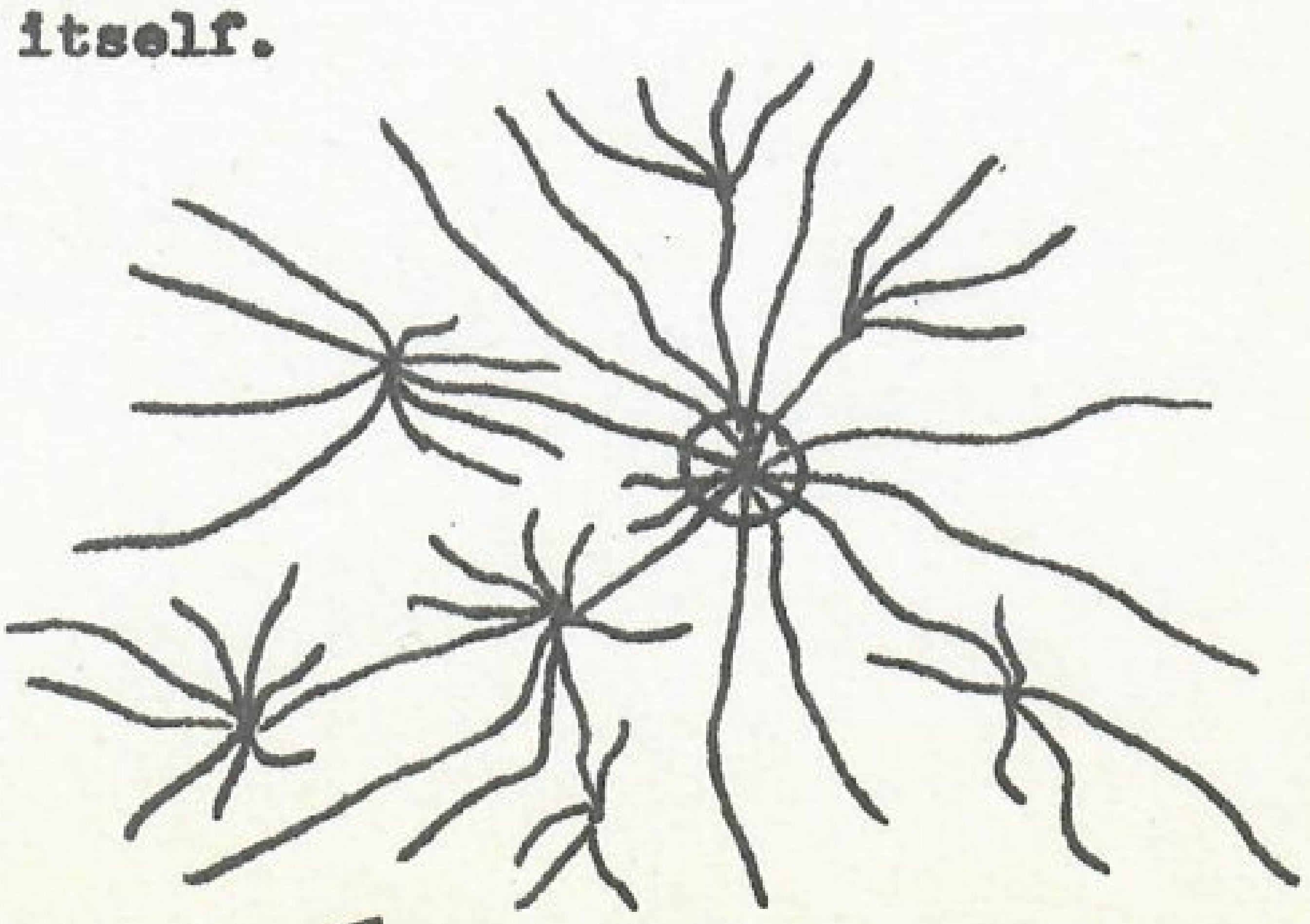

Figure 2

Figure 2 is an impressionistic attempt to depict the "flow-lines" of traffic in a mature Thünen system. The nucleus (circled) has lost some of its traffic to subordinate centers, and the clean lines of Figure 1 have become blurred, but the radial traffic pattern persists. (Radial traffic will dominate, presumably, as long as the Thünen system of zonal land-uses is not disrupted. Moving along a radius one crosses diverse land-uses which can trade with each other. Moving along a circumference one encounters only the same land-use). The direct-linkage model, in which all traffic moves without cross-hauling over a strictly radial transportation system, may be an adequate representation of this complex pattern of central and sub-ordinate trade centers.

In the following sections we will develop formally the entrepôt and direct-linkage models. They turn out to be surprisingly rich in significant theorems, deduced from relatively mild assumptions. Thünen systems turn out to be quite tractable mathematically. The decisive mathematical simplification they allow is the collapse of the location variable to a single dimension: everything of locational significance about a site is summarized in the single number giving its distance from the nucleus. (This would not be so if there were geographical irregularities, zoning constraints, or other conditions differentiating sites which are equidistant from the nucleus; but if this were so, we would no longer, in general, have a Thünen system).

We will strive for a great degree of generality in this exposition. In particular, "land uses" will not be confined to steady-state activities, but may be of an inherently dynamic character (e.g. construction or mining). In the next section, at least, very little will be demanded of the geometry of the system, and the assumptions concerning the motives and capabilities of the participants in the system will be very mild indeed. Generality of approach seems appropriate in view of the rich variety of situations which are plausible candidates for representation as Thünen systems.

Of our two special cases, the entrepôt model is definitely the more elementary, as one might expect. It is probably also the more important, as far as applications are concerned. However, even if one is only interested in the entrepôt case, there are advantages in studying direct-linkage as well. Many of the theorems that hold for the entrepôt case carry over into direct-linkage as well (though the proofs are generally more complicated), and a knowledge of these gives us deeper insight into what depends on what.

There is also a more subtle connection between the two cases. By an artifice, it is possible to reduce the entrepôt model to a special case of the direct-linkage model. This is done as follows. Suppose one had a direct-linkage system with the property that no resource appears both as an input to a certain land-use present in the system, and as an output from another land-use present in the system. (Thus, every resource flowing in the system can be classified unequivocally as an output, or as an input). Every output must then flow to the nucleus, since no other site will absorb it; conversely, every input at a site must flow from the nucleus, since there is no other site to produce it. This flow pattern apes the entrepôt case, where one also has only flows to and from the nucleus. This fact is the key to the reduction. Now take any entrepôt system. We construct an"equivalent"direct-linkage system as follows. All land-uses and resources are the same <u>except</u> that the same resource outflowing and inflowing in the entrepôt system is taken to be two different resources in the direct-linkage system. The latter has the property discussed above, that no resource is both an input and an output. It is clear that the operational properties of the two systems are identical, and the reduction has therefore been accomplished.

This result gives us the "meta-economic" theorem that any result proved for all direct-linkage systems must hold a fortiori for all entrepôt systems. We will make use of this fact in the sequel by sometimes proving theorems for direct-linkage systems and omitting a special proof for the entrepôt case.

4.2. <u>Land Uses and Land Values</u>

The model to be presented here is developed in the framework of the real estate market discussed in Section 3.1, the site-substitution principle of Section 3.2, and the concept of regional "weight" defined in Section 3.3. Until further notice we shall be dealing with entrepôt systems only.

A <u>land use</u> is simply a two-dimensional activity, spread out over a finite area of the Earth's surface. In the last chapter, this "spread" was of negligible importance, and we assumed it away to get Weberian activities, but for Thünen systems it is crucial. Weberian activities are located at a scattered set of points, and do not occupy significant quantities of land. In Thünen systems--especially in the entrepôt case--there is intense competition to get close to the nucleus, and it is precisely the finite spread of activities that prevents them from all jumping right on top of the nucleus and disperses them to a greater or lesser distance from the nucleus.*

* A more general theory would include the nuisance effects (negative neighborhood effects) of activities upon each other as a cause for dispersion. To include these would require a much more refined analytical apparatus than we have been using.

We assume everything starts at time zero, at which time the big real estate auction occurs; land is parceled out among potential renters, who assign land uses to the parcels coming under their control. These land uses then determine what transpires at all sites into the indefinite future (or up to some time horizon, if there is one). That is to say,all the essential decisions are made at the very beginning, and the passage of time simply unfolds these decisions in a foregone manner, with no revisions occurring, no new information accruing.*

* The system is "metastatic" in the terminology of W.S. Vickrey Metastatics and Macroeconomics (New York, Harcourt, Brace & World, 1964), p.3ff.

For simplicity we assume that time is discrete. (No great difficulty would attend the use of continuous time). The land use assigned to the (two-dimensional) site L determines a resource bundle a(L,t) to be delivered from the nucleus to the site L at time t, and a resource bundle b(L,t) to be delivered from site L to the nucleus at time t, for t = 0, 1, 2... (This usage reverses the roles of a(L,t) and b(L,t) from those of Section 3.3; the reason is that we are now focusing on the hinterland rather than the Weberian nucleus; an input to one is an output from the other, and we want a(L,t) always to stand for inputs and b(L,t) always to stand for outputs). Let us suppress "L," and write a_{it} for the quantity of resource i arriving at time t, and b_{it} for the quantity of resource i departing at time t.

We next assume that, for all i and t, the a_{it} and b_{it}, as measures on the system of sites, are sufficiently smooth to have spatial densities at all points (so that we may measure the arrival of a given commodity at a given time at a given point in, say, tons per acre). It will cause no confusion if we let the same symbols, a_{it} and b_{it}, stand for densities as well as quantities, depending on context. Until further notice we will be working with densities at a point, so that all costs, ideal weights, etc., are to be understood as "per unit area". The advantage of this procedure is that we now have a simple point location, and distances to and from this point can be defined unambiguously.

Just as in Section 3.3, we may define the weight-density of an activity in terms of the ideal densities of its resource inputs and outputs (or the weight, in terms of the ideal weights of these). The weight-density is the sum over inputs and outputs, and over all time, of the ideal densities of the resource-bundles involved. A spelling-out will help to clarify this definition. Suppose the metric has already been defined, and let p_{it} be the cost, in current dollars, for transporting a_{it} over unit (ideal) distance. p_{it}/z_t is the ideal weight of a_{it}, where z_t is an appropriate discount factor. ($z_t = (1+r_1)(1+r_2)..(1+r_t)$, where the r's are short-term discount rates). Similarly, let p'_{it} be the current cost for transporting b_{it} over unit ideal distance. Then the weight-density of the activity is $\sum_t \sum_i (p_{it}+p'_{it})/z_t$.

As we noted in Section 3.3, this formulation allows us to take account of improvements in transportation technology, which are reflected in secular declines in p and p'. Also, rush hours, for example, will be reflected in temporary rises in p and p'.

The weight-density of an activity has a simple meaning in the entrepôt case. It is the saving in total transport costs, per unit area per unit ideal distance, that results from locating the activity in question closer to the nucleus.

The once-and-for-all structure of the real estate auction means that leaseholds are perpetual; for our purposes, we need not distinguish these from outright sales. It is then appropriate to refer to the prices at which land is exchanged (or reserved for his own use by the owner) as <u>land values</u> rather than as rentals. Land values are dimensionally comparable to land-use weights: both are measured in time-zero dollars. Similarly, <u>land value-densities</u> are dimensionally comparable to weight-densities, both having the dimensions "time-zero dollars per unit area".

The geometry of the system may be summarized in a single function: the access perspective of the nucleus, $\mu(r)$, giving the total available area in the closed sphere of radius r about the nucleus, for all r. (See Section 2.5). We put no restrictions on $\mu(r)$ here, except that it must be non-negative and increasing.* Neither distance nor area need have any close

* The desirability of using "non-Euclidean" geometries in this context has been stressed by L. Wingo, Jr., <u>Transportation and Urban Land</u>, op. cit., p.75-80, and W. Alonso <u>Location and Land Use</u>, op. cit., p.130-133.

resemblance to geographical distance or area. For distances, this fact has already been discussed at length. For areas, one may wish to exclude land which is geographically unsuitable or pre-empted for public uses. There is an important conventional element in choosing which land is to be included in our system, and this fact may be utilized by the resourceful researcher as follows.

Below we shall make some rather strong homogeneity assumptions about the sites of the Thünen system. If in fact the actual sites are rather diverse in character, the assumptions can still be salvaged by restricting the system conventionally to sites which _are_ fairly homogeneous--e.g. to land which is zoned two-family residential, or to land on which four-story warehouses exist. It turns out that all the results of this section, at least, hold for such a fragmentary system. Otherwise stated, one may stratify a heterogeneous Thünen system into homogeneous strata so that the results of this section hold within strata, though not necessarily between strata.

A special problem arises with multiple-story structures. The simplest way of regarding, say, a ten-story building covering an acre is just as another land use which happens to involve using that building. A different approach regards the building as a way of increasing access perspective, by placing ten acres where there was one acre before (less than ten, if one takes account of setbacks, stairwells, elevators, etc.). The trouble with this second approach is that stories are not perfect substitutes for each other, and that, while some activities may thrive on upper stories (e.g. office activities) others do poorly (e.g. assembly-line production processes). (One might say they have different degrees of stackability}. The first approach will be used in this chapter; a simple model more akin tothe second, in the next chapter.

The remaining assumptions for our entrepôt model are those used in setting up the site-substitution principle of Section 3.2. To recapitulate briefly: all sites having the same area are assumed to be equivalent, in the sense that any land use technically feasible on one site is feasible on another, and there are no zoning laws imposing different constraints on different sites. All sites are equally available (or unavailable) to all individuals. There is no discrimination, no absolute locational preferences, and market information is perfect.

Two points call for comment. First, one may wonder how to reconcile the assumption of perfect information made here with the assertion that imperfect information is an important cause for the formation of entrepôt systems, which we made in the previous section. The answer is that the degree of "perfection" is different. Perfect market information requires only that people know the going prices and locations of all parcels. This is compatible with not knowing the locations of one's potential trading partners. Ignorance of the latter is what centralizes trade.

The second point concerns the divisibility of land uses. Suppose a given land use covers two acres. If each acre is considered to be a separate site, then by the equivalence assumption the overall land use could be separated into two pieces at nonadjacent locations without prejudice to the feasibility of either half of the original land use. The equivalence assumption implicitly denies the influence of the environment on what is technically feasible at a site, even if the "environment" is, say, the other half of a building. In other works, neighborhood effects are ignored.* The error so introduced is probably not too important if the realm of influence of neighborhood effects is very small in relation to the size of the entire Thünen system: for example, if they occur only within individually-held parcels, and these are very small parts of the whole.

This last group of assumptions leads to the site-substitution principle of Section 3.2, which states, it will be recalled, that, given one's entire location plan except for the location of one activity, one chooses the site minimizing the sum of transportation costs plus on-site costs. On-site costs here reduce to the single item of land value. (Resources originally located on the site are included in the real estate package; all other resource inputs come from the nucleus, and are recorded under transport costs). Transportation costs, by our construction, are the product of the weight of the land use by the (ideal) distance between site and nucleus.

It is instructive to compare the form assumed by the site-substitution principle here with the form it assumes for Weberian activities. Land value is negligible for Weberian activities, but not for land uses in the Thünen context. But this complication is more than compensated for by the great simplicity of

* In this case, positive neighborhood effects (usually called "indivisibilities"--a less perspicuous title). We have already mentioned that an adequate treatment of neighborhood effects would require a more profound set of concepts than we have been using.

transportation costs in the entrepôt case. In general, for Weberian activities resource flows come and go in any number of different directions, precluding any simple relation between cost and location. In the entrepôt case, all flows come and go between the site and a single point--the nucleus--and transportation costs are simply proportional to the distance between site and nucleus. This fact is the key to the great simplicity of the entrepôt case, and the explanation for the rather powerful theorems which can be derived in this case.

We will use the site-substitution principle in density form. For our entrepôt case, this states that one chooses the location minimizing the sum of land value-density plus the product of weight-density by distance to nucleus. Only in this form does "distance" have an unambiguous value.*

* The meaning and validation of the site-substitution principle in density form involves the comparison, between competing sites, of a sequence of regions about these sites, the sequence shrinking to zero as a limit in diameter and area. It will suffice to think of "spheres", of area ϵ, about the competing points, where ϵ is "very small". One then has to find the optimal sphere in which to locate the land use in question, and this sphere is determined by the site-substitution principle in density form, as applied to the centers of the respective spheres.

So much for preliminaries. We now come to the substantive results. The first is a general ordering principle for land uses. The following ones refer to the structure of land values.

__Theorem 1__: In an entrepôt system, in which the site-substitution principle (in density form) is satisfied, land use weight-densities are a non-increasing function of distance from the nucleus (except possibly at a set of distances of Lebesgue-measure zero, for which several weight-densities may attach to the same distance).

__Proof__: Let us compare two points, L' and L'', at distances r' and r'' from the nucleus; let v' and v'' be the land value-densities at the two points; the land uses being run at these points are A' and A'', having weight-densities w' and w''; assume that $r' < r''$.

The user of point L' preferred this location for A' over the feasible location at L''. Total land value plus transportation costs are $v' + w'r'$ at L', and would be $v'' + w'r''$ if he chose to run A' at L''. By the site-substitution principle we must then have

1) $$v' + w'r' \leqq v'' + w'r''.$$

Analogous reasoning for the user of point L'' leads to the inequality

2) $$v'' + w''r'' \leqq v' + w''r'.$$

Adding these inequalities and canceling yields $(r''-r')(w''-w') \leqq 0$, and, upon dividing by the positive number $(r''-r')$, we get $w' \geqq w''$. This proves "non-increasingness" when the two points are at different distances from the nucleus. When the two points are at the same distance from the nucleus, however, the last step breaks down, and no conclusion can be drawn: there may land uses having different weights at a given distance. But it may be shown from the "non-increasing" property that such anomalies can occur at most at a countable number of points. (One cannot place an uncountable number of non-overlapping intervals on the real line). This proves the parenthetical qualification. QED

The main conclusion of Theorem 1 will be called the __weight-falloff condition__: if $r' < r''$, then $w' \geqq w''$.

Figure 3 portrays some restrictions imposed by, and possibilities compatible with, Theorem 1. Distance r_1 is one of the anomalous distances, and activities with weights ranging from w_4 to w_5 occur at that distance. Between r_2 and r_3 is a flat stretch in which all land uses have weight-density w_3. At r_4 there is a discontinuous drop in weight-density from w_2 to w_1, as might occur at a sharp break in land uses.

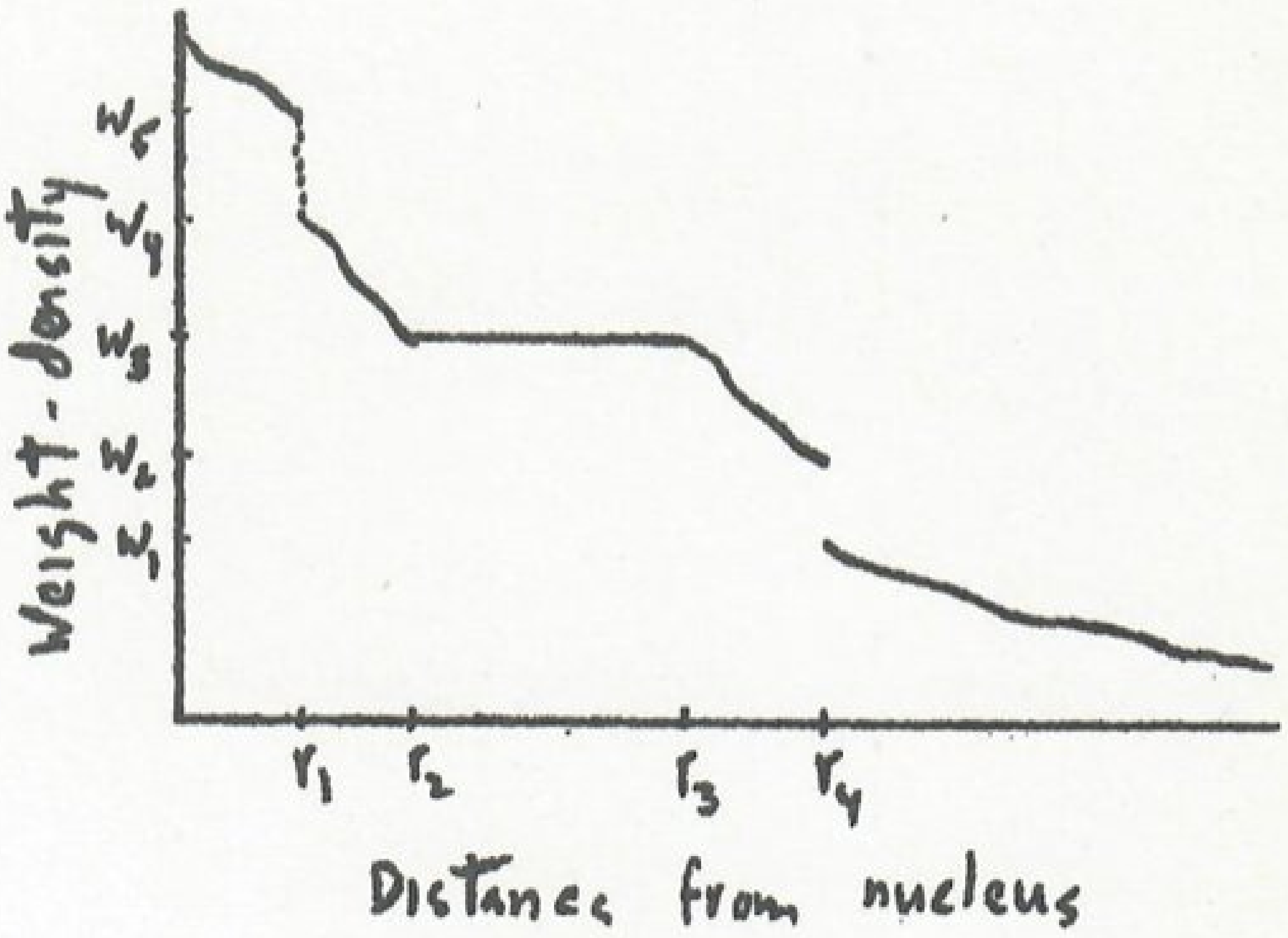

Figure 3

The anomalous case might occur in the following situation. Suppose, because of freight absorbsion or some other quirk in the transportation system, that the distance from the nucleus were equal over a finite zone of the entrepôt system. Then it is possible that different land uses with diverse weight-densities would co-locate in this zone. Whether this case has any practical significance is a moot question.

The reader will have noticed that no use was made of the Thünen assumption, that land uses at a given distance are uniform, in Theorem 1. This was done to see to what extent the Thünen pattern could itself be derived from more fundamental assumptions. Theorem 1 indicates that we can come close to, but not quite reach, the Thünen pattern: except for the qualification, any two land uses at the same distance must have the same weight, though they need not be identical. If one makes the Thünen assumption outright, then of course the qualification to Theorem 1 may be dropped. Occurrences such as that at r_1 in Figure 3 cannot happen, and we may simply state that weight-density is a non-increasing function of distance from the nucleus.

Conclusions similar to that of Theorem 1 are quite common in the literature, though they are usually phrased in terms of "intensity of cultivation" or "density of occupancy", rather than weight-densities.* The fundamental contribution of Theorem 1 is in the weakening of assumptions. The remarkable generality of the conditions under which weight-density decreases with distance seems never to have been realized. We recapitulate as follows.

1) The land uses may be quite complex in form, involving multiple products, and inputs and outputs in an arbitrary time pattern. The concept of weight-density must of course be correspondingly more sophisticated than, say, tons per acre of output.

2) The land uses may be completely variegated by type. Almost all treatments have been of agricultural and/or residential land uses, but we may include manufacturing, commercial, office, religious, and other land uses, all bidding against each other in the same omnibus real estate market.

3) We need not assume any simplicity or homogeneity in the preferences of the various land users. Some may be profit maximizers, others not. Tastes in residences and consumer goods may differ widely. The only requirement, besides the ones of no discrimination and no absolute location preference which have been discussed, is the trivial one that, ceteris paribus, more money is preferred to less. (See Section 3.2).

* Cf. E.S. Dunn The Location of Agricultural Production (Gainesville, University of Florida Press, 1954); Wingo Transportation and Urban Land, op. cit.; Alonso Location and Land Use, op. cit.

3½) Nothing need be assumed concerning the incomes of individuals or their technical capabilities.

4) Individuals need not be simply-located. They may have multiple residences, or multiple businesses, or several of each. The discussion of Section 3.2 indicates that such multiple-location has no effect on the validity of the site-substitution principle.

5) Aside from the strong entrepôt assumption, the geometrical assumptions are so weak as to be trivial.

The proof of Theorem 1 is of independent interest, due to its great simplicity. It will be shown below that the same line of argument gives substantial information concerning the structure of land values.

We now go on to applications of Theorem 1. Theorem 1 cannot tell us which land uses, out of all those technically feasible, will actually be put into practice on some site or other. Nor can it tell us how much territóry these land uses will cover. If we are already given the fact that two land uses are being put into practice somewhere in the system, Theorem 1 tells us which will be more central and which more peripheral; that is, it orders land uses by their distances from the nucleus. (The only ambiguity results from the borderline case when two land uses have the same weight-density.)

A fairly diverse number of questions may be raised which are in a form suitable for the application of Theorem 1. For example, why do buildings tend to get taller as we approach the center? Why does land speculation occur largely on the periphery of a Thünen system? Why are office-buildings usually quite central? What kind of manufacturing activities tend to de-centralize? Why does population density fall off with distance, and does it always? What determines the distribution of people by income? By automobile ownership?

One can, of course, find special-purpose explanations for each of these questions, and others of the same form. But Theorem 1 permits a unified approach to all of them. There are of course caveats to be observed. The entrepôt assumption, and the premises underlying the site-substitution principle, may be so badly off the mark that Theorem 1 becomes useless for explanatory purposes. Also, these questions are not specific enough for Theorem 1 to make an unhedged prediction; consequently, only a statement of "tendencies" can bé expected.

Let us first apply Theorem 1 to multiple-story structures. Multiple-story structures are, in effect, several land uses stacked one atop the other, and surrounded and held in place by some edifice. Since they are all over the same site, it follows that the weight-density of the whole is the sum of the weight-densities of the separate land uses, plus the weight-density added by the original construction (coming and going of construction workers and equipment, inflows of cement, steel, glass, etc.) We may expect, then, a fair positive correlation between number of stories and weight-density (not perfect correlation, since a smaller number of "heavy" one-story land uses can outweigh a larger number of "light" ones). From Theorem 1 it follows that the "denser" high-rise structures should be found toward the center and the low-rise structures (and the open land) should be found toward the periphery.

Next, <u>land speculation</u>. By this we mean the deliberate delay in initiating a land use on a given parcel, for any reason whatsoever. Now a delayed land use can be thought of as simply a different land use, related to the original by having all inflows and outflows displaced forward in time, say by T periods, and having the first T periods blank. This may be called the T-displacement of the original land use. How does the weight-density of these various land-uses depend on T? As a rule, the weight-density should decline with increasing T, due to a combination of discounting and secular improvements in transportation. (If p_{it}/z_t is a decreasing function of t, for all i, it is easy to show that density must decline with displacement). Now apply Theorem 1. If both a original land use, and a displacement of it, are applied somewhere in the entrepôt system, the displacement, being lighter, will lie beyond the original. If a whole sequence of displacements co-exist, they will be ranged in order of increasing displacement outward toward the periphery. This gives a rough picture of the usual pattern of land speculation, which concentrates in the suburbs.

This whole argument can also be turned around. Suppose we consider two land uses, a "heavy" one and a "light" one. It may be that a displacement of the "heavy" land use still remains "heavier" than the original "light" one. Suppose that, as it turns out, the original "light" use, and the displaced "heavy" use, co-exist in the entrepôt system. At a certain moment in time, the "light" use will be already under way, while the displaced "heavy" use will still be in the offing, the land destined for it sitting silent, waiting its appointed hour. By Theorem 1 the displaced activity will be located more centrally than the "light" one, and cross-sectionally the system will resemble Figure 4:

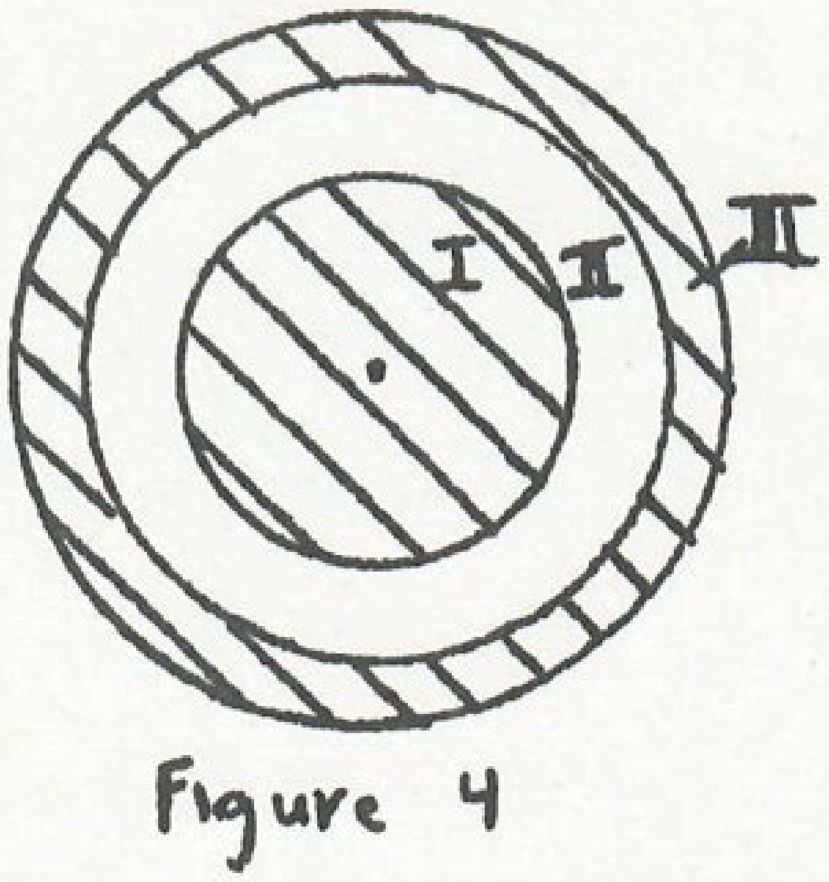

Figure 4

Zone III has the "light" use on it, and, together with zoneI, it is already humming with activity; while zone II, which is reserved for the displaced use, is still inactive. This phenomenon is called leapfrogging. It is interesting that it can arise in a perfect market, so that the usual explanations in terms of diverse expectations and imperfect information among landowners need not be invoked.#

Cf. W.R. Thompson A Preface to Urban Economics, op. cit., p.326f. This is not to deny that these factors also play a role, perhaps the major role.

Spatial differentiation by major type of land use--e.g. manufacturing, residential, agricultural--may also be treated by Theorem 1. The characteristic range of densities of these major types places them in a characteristic position in the sequence from centrality to peripherality. Thus, the low density of agriculture makes it peripheral. Manufacturing offers an instructive example, because the land uses embraced under this heading have such an extreme range of densities. The relatively central manufacturing activities are those which lend themselves readily to stacking in multiple-story structures, while long low-slung plants are characteristically suburban.#

Curiously, one gets the impression that so-called "light" manufacturing is more typically stacked in multiple-story buildings than "heavy" manufacturing, and that the former therefore tends to have the heavier weight-density. A plant in the garment district is denser than a sprawling steel mill, for example.

The remaining applications of Theorem 1 to be discussed here concern residential land uses. The most important contribution to the weight of a residential land use will usually be the trips people make to and from the residence. The weight contributed by a given trip depends on who is making the trip, when it is being made, and the mode of transportation, among other factors. The summation over all trips by a given person determines his total direct contribution to the weight of the land use. (He will make indirect contributions in the form of extra inflows of consumer goods, and in other ways). Finally, the summation over all persons gives the total (direct) contribution by person-trips to residential weight. To convert this to weight-density, one divides by the area of the site occupied by the residential land use, the area including homesite, grounds, garage space if any, etc.

Very schematically, we may analyze the weight-density contributed by trips into the product of three factors: (1) cost per unit (ideal) distance per trip; (2) frequency of trips per person; (3) persons per unit area. If the first two factors are fixed, then weight-density will vary directly with population density, and Theorem 1 then predicts that population density should in general decline with increasing distance from the nucleus. But with diverse population types, weight-density need not vary monotonically with population density, and the latter, therefore, may behave irregularly over space.* For example, retired people will probably make fewer trips per person on the average than employed people; also, their costs per unit distance

* A possibility noted by Wingo, op. cit., p. 100.

will probably be lower, since they do not forego income in spending time traveling. Thus the contribution to weight of each retired person will be light, and as a result we may find densely populated retirement communities at a distance that would sustain only a very sparse employed population density.

The relation between income and residential location has been the subject of a sizable literature. How does average income of residents vary with distance from the center?* According to Theorem 1, the answer to this question depends on how weight-density of residence varies with income. Suppose we combine the first two factors--transport-cost per unit distance, and trip frequency --to give transport cost incurred per unit distance per unit time (per person). For short, let us call this cost-intensity. Let us also take the reciprocal of the third factor, to get area per person. We then have the following theorem.

* The empirical data on this question are quite mixed. See L.F. Schnore "The Socio-economic Status of Cities and Suburbs" American Sociological Review 28:76-85 February, 1963, and "On the Spatial Structure of Cities in the Two Americas", Chapter 10 of The Study of Urbanization (P.M. Hauser and L.F. Schnore, eds., New York, Wiley, 1965). For theoretical approaches see

Wingo, op. cit., p.95-100; Alonso, op. cit., p.106-109; G.S. Becker "A Theory of the Allocation of Time", op. cit., p.511f.

Theorem 2: Suppose (1) a Thünen entrepôt system satisfies the site-substitution principle; (2) only personal trips contribute to the weight-density of residential land uses; (3) there is a population characteristic, y (e.g. income), such that the choices of cost-intensity, c, and area per person, a, are both differentiable functions of y. Then distance from the nucleus increases with y if the elasticity of a exceeds the elasticity of c (both with respect to y), and decreases with y in the opposite case.

Proof: The weight-density of a residential land use is proportional to c/a. This decreases with rising y if the elasticity of a exceeds the elasticity of c, and increases in the opposite case. Now apply Theorem 1. QED

The "elasticity of area per person with respect to income" is self-explanatory, but the meaning of "elasticity of cost-intensity with respect to income" requires some discussion. As income rises, the time spent in traveling becomes more expensive, as a rule, in that it probably represents greater foregone income. If time-delay is the only element of transportation cost, and no other adjustments are made, one might assume that cost-intensity is proportional to income.*

* This is the procedure of Gary Becker, ibid., in his very succinct model. The elasticity of c is then equal to one, so that distance rises with income if the elasticity of a exceeds one. This is also Becker's result, of course--obtained, incidentally, by a completely different route than the proof of Theorem 2.

There are usually several factors which make the rise in cost-intensity less than proportional to the rise in income, however. (1) Larger incomes--at least past a certain point--have a greater fraction of unearned income than smaller, and unearned income does not make time more expensive; (2) monetary transportation costs do not vary with income (for the same mode of transportation); (3) in response to rising time-costs with income, there may be a cut in trip frequencies (e.g. by making more overnight trips, or by cutting down on recreational trips); (4) in response to rising time-costs

one may choose faster means of transportation (e.g. if congestion is not too severe, private automobile, which avoids waiting for public transportation; or, one may simply drive faster); (5) over and above the substitution effect of point(4), a pure income effect leads wealthier people to own more and better automobiles; this makes them generally "lighter" in ideal weight and so reduces transport costs.*

* But also leads to a more than compensating increase in trip frequencies (see Section 1.6.)

If the overall effect of these factors is to make the elasticity of cost-intensity less than one, then the elasticity of demand for space can also be less than one without implying that distance from the nucleus declines with increasing income.

It is rather surprising that this entire discussion could be carried on without having to refer to land values at all.

We now get back on the main track, and see what can be deduced about the pattern of land values in our entrepôt system. It will be shown that (1) land value is a function of distance from the nucleus; (2) this function is continuous; (3) this function decreases up to any distance at which land is not permanently vacant; (4) if weight-density is continuous at a certain distance, thenthe derivative of land value exists there, and it equals minus weight-density; (5) land value is a convex function of distance.

These statements will be proven in a sequence of theorems. The basis on which they all rest are the inequalities (1) and (2) used in the proof of Theorem 1; we repeat these here for convenience:

1) $v' + w'r' \leqq v'' + w'r''$; (2) $v'' + w''r'' \leqq v' + w''r'$,

where v' and v'' are land value-densities, w' and w'' are weight-densities, and r' and r'' are distances, for the points L' and L'', respectively, with land uses A' and A''.

Theorem 3: Land value-density is a function of distance from the nucleus.

Proof: We must show that if $r' = r''$, then $v' = v''$. Assuming the first equality and substituting in (1) and (2), we get $v' \leqq v''$, and $v'' \leqq v'$; therefore $v' = v''$. QED

This result should be contrasted with Theorem 1, where it could not be proved that equi-distant points carried equi-dense land uses.

<u>Theorem 4</u>: Land value-density decreases with distance up to any point at which land is not permanently vacant (i.e. up to any point at which weight-density is still positive).

<u>Proof</u>: Assume that $r' < r''$. From (2) we get $v'-v'' \geqq w''(r''-r') > 0$.
QED

<u>Theorem 5</u>: Land value-density is a continuous function of distance.

<u>Proof</u>: By a re-arrangement of (1) and (2) we get

3) $$w''(r''-r') \leqq v'-v'' \leqq w'(r''-r').$$

Now let, say, L'' be a variable point and approach L' as a limit. As $r'' \rightarrow r'$, both ends of this double inequality go to zero, and so $v'' \rightarrow v'$. QED

Again, both these theorems offer instructive contrasts with Theorem 1. It cannot be proved that weight-density decreases with distance, but merely that it does not increase. Also, weight-density need not be a continuous function of distance.

<u>Theorem 6</u>: If weight-density is continuous at r', then the derivative of value-density exists at that distance, and equals $-w'$.

<u>Proof</u>: Suppose weight-density is continuous at r', and that $r' < r''$. A re-arrangement of the inequalities (3) yields

4) $$-w'' \geqq (v''-v')/(r''-r') \geqq -w'.$$

As r'' approaches r' from above, $-w'' \rightarrow -w'$, by continuity; therefore, $(v''-v')/(r''-r') \rightarrow -w'$; for $r'' < r'$, merely reverse the signs in (4) and repeat the argument. QED

(Theorem 6 can be strengthened slightly to read: at any distance, the (right-hand, left-hand) derivative of value-density exists, and equals minus the (right-hand, left-hand) limit of weight-density at that distance.)

<u>Theorem 7</u>: Value-density is a convex function of distance (i.e., if three distances r', r'', and r''', ane such that $r'' = \lambda r' + (1-\lambda)r'''$, where $0 < \lambda < 1$, then $v'' \leqq \lambda v' + (1-\lambda)v'''$.)

<u>Proof</u>: Assume the premise; we have $v'' + w''r'' \leqq v' + w''r'$, and also $v'' + w''r'' \leqq v''' + w''r'''$; multiply the first inequality by λ, the second by $(1-\lambda)$, and add, to get

$$v'' + w''r'' \leqq \lambda v' + (1-\lambda)v''' + w'[\lambda r' + (1-\lambda)r'''].$$

From the premise, the terms containing the r's drop out, leaving

$$v'' \leqq \lambda v' + (1-\lambda)v'''.$$

QED

We portray a typical value-density curve in Figure 5 compatible with Theorems 3 through 7, just as we did for weight-density in Figure 3. The value-density curve is much more restricted, since it must be single-valued, strictly decreasing, continuous and convex. Before r_1, and between r_2 and r_3, it is strictly convex. Between r_1 and r_2, and after r_3, it is linear. There are "kinks" at r_1, r_2, and r_3. The linear stretches correspond to flat stretches in the weight-density curve, and the kinks correspond to discontinuous drops in that curve.

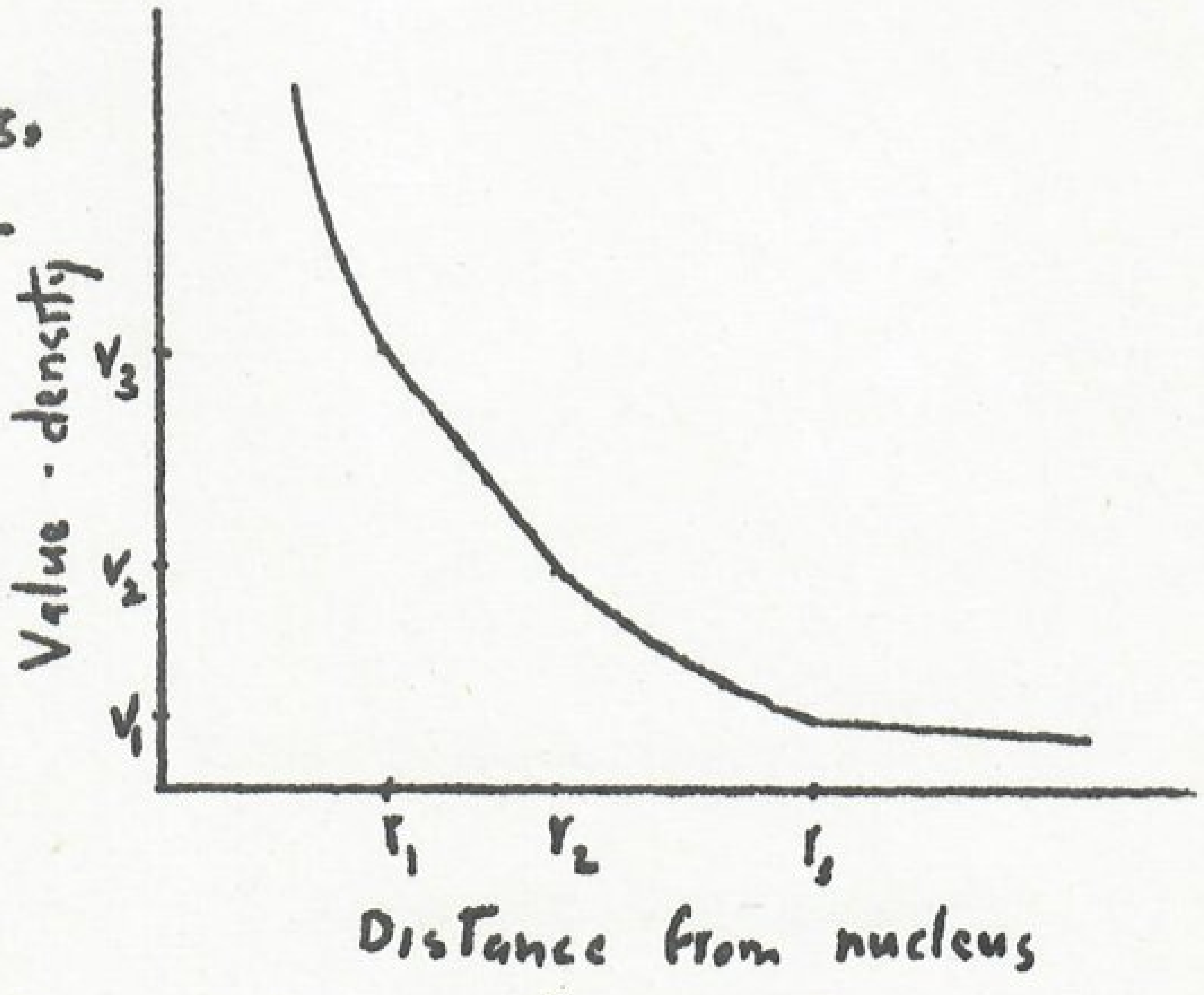

Figure 5

These results must be interpreted with some care. All land values are as of time zero, when the landscape is uniform in geographical characteristics. As diverse uses are put into operation, this pristine uniformity will, in general, be destroyed. The following interesting possibility then arises. Suppose some inner ring has a land use with a quick "pay-off" that eventually leaves the land in a dilapidated condition, while some outer ring has a land use with a slow "pay-off" that leaves useful structures. We may then find at some future time that the neat cross-sectional pattern of Figure 5 has been undermined, and that land values actually rise over some distance ranges. This possibility cannot be laid to rest without further assumptions. A simple assumption that would do the trick is that no land uses modify the technical characteristics of a site (e.g. no construction or mining occurs). The decaying cores of cities may show that the possibility raised here is of more than theoretical interest.

Concerning Theorems 3 through 7, the same general comments apply as were made concerning Theorem 1. Similar results have been obtained in the literature. Our treatment is more rigorous and simple, and, most important, our results hold under far more general conditions than have been assumed. (Cur assumptions are the same as for Theorem 1).

Figure 5 gives the pattern of land values in terms of ideal distances. In terms of geographical distances the pattern may be rather scrambled. For example, our results are quite compatible with there being a region of high land values encircling a suburban railway station, since such points may be functionally close to the nucleus. There is, however, at least one case in which the pattern of Figure 5 carries over to geographical distances, and which has practical significance:

<u>Theorem 8</u>: Suppose ideal distance from nucleus is an increasing concave function of geographical distance from nucleus, and that --as in Figure 5--value-density is a decreasing convex function of ideal distance. Then value-density is a decreasing convex function of geographical distance.

<u>Proof</u>: Let g be the function mapping geographical into ideal distance, and v the function mapping ideal distance into value-density. We are to show that the composition $v \circ g$ is decreasing and convex.

$x' > x'' \Rightarrow g(x') > g(x'') \Rightarrow v(g(x')) < v(g(x''))$, so $v \circ g$ is decreasing.

Suppose $x'' = \lambda x' + (1-\lambda)x'''$, where $0 < \lambda < 1$. Then $g(x'') \geqq \lambda g(x') + (1-\lambda) g(x''')$, by concavity of g. $\therefore$

$$v(g(x'')) \leqq v(\lambda g(x') + (1-\lambda) g(x''')) \leqq \lambda v(g(x')) + (1-\lambda) v(g(x'''))$$

The first inequality on this line follows from v being decreasing, the second from v being convex. Together they imply that $v \circ g$ is a convex function. QED

The concavity condition on g means that, the more distance one has already covered, the less costly it is to go an extra unit distance. Many freight tariffs have this property, but it is probably false for car travel, because of increasing fatigue and scarcity of time.*

* Cf. E.M. Hoover <u>The Location of Economic Activity</u>, op. cit., p.75f. Hoover has the essentials of Theorem 8, but states the condition on g ambiguously: as less-than-proportional response, and also

as diminishing marginal response, which is correct. The former condition (which is diminishing average response) is not sufficient to derive the conclusion of Theorem 8.

4.3 Direct-linkage Models

It will be recalled that direct-linkage models are Thünen systems in which sites may ship directly to each other, rather than having to trade only with the nucleus. It was pointed out in Section 4.1 that an entrepôt model was equivalent in a certain sense to a special kind of direct-linkage model, and the results of this section will, therefore, have as corollaries corresponding results concerning entrepôt models.

Direct-linkage models retain the indispensable "one-dimensional" feature of Thünen systems. But the concepts of "weight" and "weight-density", which played so fundamental a role in the last section, are fairly useless here. The reason is that, while some of the inputs and outputs of a land use may travel inward toward the nucleus, others may travel outward. The net inward pull on the land use is given by the weight of the former components *minus* the weight of the latter. In the entrepôt case everything pulls inward; in the direct-linkage case there are opposing pulls, and--this is the crucial point--which components pull in which directions depends on the rest of the system: The effective net weight of the land use is no longer an intrinsic property of the use alone.

To compensate for the dropping of the entrepôt assumption, we will, later on, strengthen the geometrical assumptions, and assume that access perspective is a power function. This case, which includes the Euclidean line, plane, and space, is less restrictive than appears at first glance, and yields some strong and interesting results.

The previous section was framed in a social equilibrium setting, with very little being assumed about the preferences or capabilities of the individual participants. The present section, by contrast, will concern itself with the optimal assignment of land uses with respect to a single overall preference order, which m makes it more appropriate for a regional planner, or for a farmer laying out his fields. The preference order will be of the cost-minimization type--akin to that of the headquarter location models of Chapter 3--or, more generally, of the income-maximization type. (The meaning of these phrases in the present context will be spelled out below).

One can sometimes show that a social equilibrium system behaves as if it were maximizing a single utility function. If a direct-linkage model can be shown to have this property, then the results of this section will be applicable to it. An interesting example will be given later.

We will need a more detailed discussion of land uses and traffic flows in this section than in the last. Let A(r) be a function giving, for every distance from the nucleus, r, the land use, A, running at that distance ("A" for "activity"). By our Thünen assumption, every site at distance r carries the same land use, so A(r) is single-valued. A(r) determines two functions, a(r) and b(r), giving the input and output densities, respectively, at distance r. Both a and b are vectors of rather big size, ranging over all commodity-time combinations: $a = \{a_{it}\}$; $b = \{b_{it}\}$. All the components of a and b have the same dimensions, viz., (ideal) weight per unit area.

The function q(r) gives the traffic flow across the rim of the"sphere"of radius r about the nucleus.* We adopt the convention that an inflow toward the nucleus is positive, and an outflow negative. q is a big vector of the same conformation as a or b, so that q(r) gives, for each distance, the entire history of flows of all commodities: $q \equiv \{q_{it}\}$ All the components of q have the dimension of (ideal) weight; thus $q_{it}(r)$ is the ideal weight of commodity i at discrete time t flowing through the rim of the sphere of radius r.

(In the entrepôt case, q would have to be twice the size of a or b, since it would have to account for flows in both directions for each commodity and time. Under direct-linkage, there is no cross-hauling, so, for each commodity and time, either the inward or the outward flow (or both) is zero, and we may adopt our simple sign convention).

The functions a(r), b(r), and q(r) are connected by a fundamental material balance relation:

Theorem 1: $q(r) = \int_r^\infty (b(r) - a(r))\, d\mu(r)$

where $\mu(r)$ is the access perspective of the nucleus, and we use the Stieltjes integral notation.*

* An ambiguity in the definition of q(r) arises if there is a finite blob of area at a given distance. We adopt the convention that q(r) is continuous from the left at such distances. For consistency, the integrals below are to be understood as defined on closed half-lines.

Proof: For any commodity i and time t, $b_{it}(r) - a_{it}(r)$ is the net output of that commodity at that time per unit area at distance r. The differential of access perspective gives the area available for the activity A(r) over a small increment of distance, and the integral of their product from r to infinity gives the total net production of i at t in the entire Thünen system outside the (open) sphere of radius r about the nucleus.

Part of the definition of direct-linkage is that trade with the outside world occurs only through the nucleus. This being so, a net production surplus has nowhere to go but through the rim of the sphere toward the nucleus. Conversely, a net production deficit has nowhere to be made up from except by an outflow through the rim of the sphere away from the nucleus.

This is true for each i,t combination, and all these results together may be summarized in a vector integral. Theorem 1 then follows from our sign convention that inflows are positive and outflows negative. QED

It is instructive to contrast Theorem 1 with the result in the entrepôt case. Inflôws and outflows must be handled separately. The vector of inflows toward the nucleus at distance r is given by $\int_r^\infty b(r)\,d\mu(r)$ and the vector of outflows away from the nucleus is given by $\int_r^\infty a(r)\,d\mu(r)$Comparison with Theorem 1

* We use the standard convention that the integral of a vector function is the vector whose components are the integrals of the component functions.

shows that, for the same land use assignment, direct-linkage traffic flows are simply the algebraic difference of oppositely-directed entrepôt traffic flows.

The total net volume of production for the entire system equals $\int_0^\infty (b(r) - a(r))\, d\mu(r)$, and this in turn equals q(0), the net traffic flow into the nucleus. (This may then be exported, or, if negative, made up by imports; or it may be used in a Weberian activity at the nucleus. In any case we are not concerned with this aspect, only with land uses in the field).

Access perspective, $\mu(r)$, is a monotone increasing function. It therefore has an inverse, which may be written $r(\mu)$. (Wherever $\mu(r)$ has a jump discontinuity the inverse function will have a gap in its domain. We fill in this gap by taking the value of r at which the jump occurs as the value of the function in this interval (See Figure 6). This extended inverse function will also be denoted by $r(\mu)$ and is the one that will be used below. It is clearly non-decreasing.)

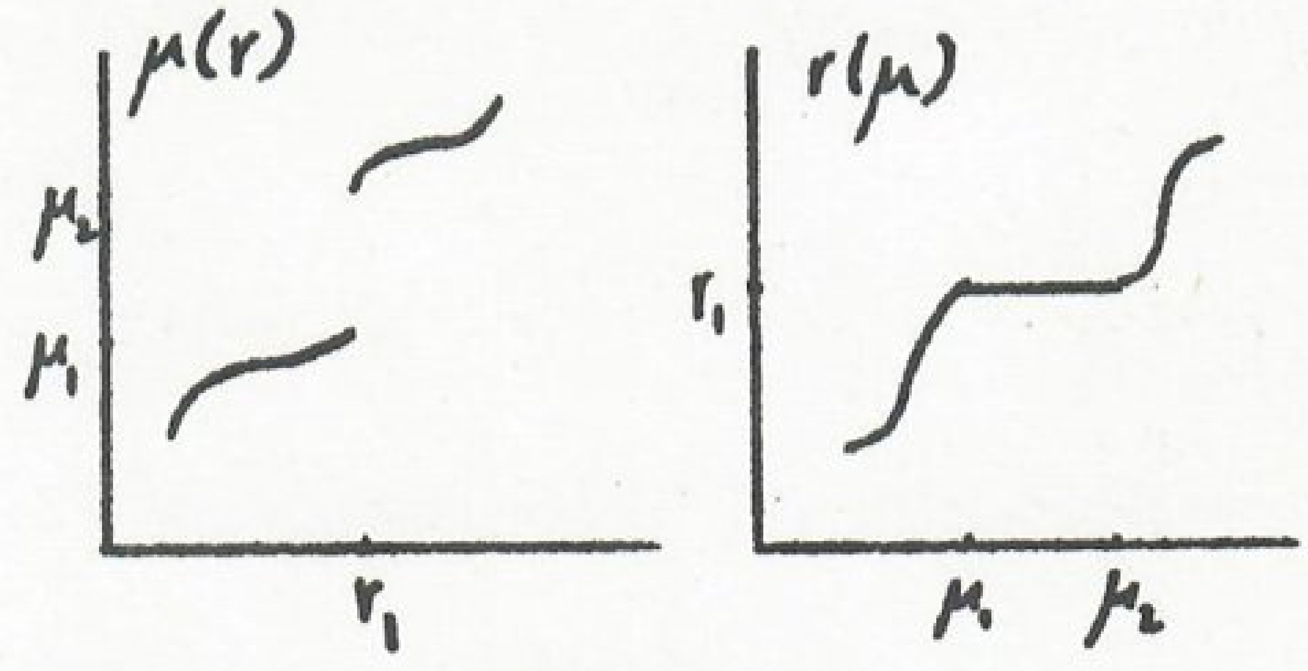

Figure 6

Input and output density, and traffic flow--a, b, and q--are all functions of r. Since r in turn is a function of μ, these are also functions of μ, and may be written as $a(\mu)$, $b(\mu)$, and $q(\mu)$. In words, $a(\mu)$ is the input density of the land use in operation at the rim of the"sphere" whose area is μ; similarly for $b(\mu)$; $q(\mu)$ is the traffic flow through the rim of this sphere. These are rather unusual functions of course, but it turns out for much of this section that μ is a more natural independent variable than r.

In terms of μ as independent variable, Theorem 1 how reads:

$$5) \qquad q(\mu) = \int_{\mu}^{\infty} (b(\mu) - a(\mu))\, d\mu\,,$$

where we now have only an ordinary Riemann integral to deal with.

Suppose we have two land use assignments, $A'(\mu)$ and $A''(\mu)$--using μ as the independent variable--which are related in the following way: $A''(c\mu) = A'(\mu)$ for all μ, where c is some positive constant. If c is greater than one, A" is a kind of spread-out repetition of A': whatever land use A' assigns to a given distance will be assigned by A" to some greater distance.*

The input and output densities and traffic flows generated by A' are denoted by a', b', and q', respectively; similarly, a", b", and q" are generated by A". We then obtain

<u>Lemma 1</u>: If $A''(c\mu) = A'(\mu)$, then $q''(c\mu) = cq'(\mu)$.

<u>Proof</u>: $q''(c\mu) = \int_{c\mu}^{\infty} (b''(x) - a''(x))\,dx = c\int_{\mu}^{\infty} (b''(cy) - a''(cy))\,dy$ $= c\int_{\mu}^{\infty} (b'(y) - a'(y))\,dy = cq'(\mu)$. The first equality is from Theorem 1, the second from the substitution $cy = x$, the third from $A''(c\mu) = A'(\mu)$, and the fourth from Theorem 1 again.

QED

* Under certain conditions it may be impossible for both A' and A" to be Thünen systems, since several land uses may have to be assigned to sites at the same distance from the nucleus.

Note in particular that $q''(0) = cq'(0)$. That is to say, the net output of the entire system has been multiplied by a factor c in going from the assignment A' to A". This leads us to the definition: A" is the c-fold expansion of A' when they are related as in the premise of Lemma 1.

We now specialize to access perspectives of the form $\mu(r) = (r/\theta)^D$, where D and θ are positive numbers. This is the class of homogeneous access perspectives. For the Euclidean line, plane and space, we get $D=1$, 2, and 3, respectively, so it is natural to call the parameter D the dimensionality of the system.# This term should not be taken too literally, as we might easily find systems on the surface of the Earth for which $D=2$ does not give the best fit. (For example, a riparian nation such as Egypt might best be thought of as a one-dimensional Thünen system). Nor need D be confined to integer values.

What does a c-fold expansion look like when the access perspective is homogeneous? We have to translate the condition $A''(c\mu) = A'(\mu)$ back into r as the independent variable. Since $r(c\mu) = \theta(c\mu)^{\frac{1}{D}} = c^{\frac{1}{D}}(\theta\mu^{\frac{1}{D}}) = c^{\frac{1}{D}}r(\mu)$, it follows that $A''(c^{1/D}r) = A'(r)$ when land use is expressed as a function of r. That is, the land uses of A' are spread out in such a way that

# For these three cases the "spheres" about the nucleus are intervals, circular discs, and solid spheres proper, respectively; the measures μ are lengths, areas, and volumes, respectively. We have been referring to these all indiscriminately as "areas", because of the pre-dominant importance of the two-dimensional case.

all distances are multiplied proportionally, by the factor $c^{1/D}$. It can be shown that the homogeneous access perspectives are the only ones for which this is true. For all others, a c-fold expansion multiplies distances in a more or less irregular manner. This simple property of homogeneous access perspectives is the key to the strong results which follow.

We next wish to consider the total transportation costs incurred over the whole Thünen system. The traffic flow of commodity i at time t is measured in ideal weight, so the cost incurred per unit distance is given by $|q_{it}|$ --the absolute value, since the same cost is incurred for inflow as for outflow. It follows that the cost incurred for commodity i at time t is given by

6) $\int_0^\infty |q_{it}(r)|\,dr,$ using r as the independent variable.

Total transportation costs is then the summation of (6) over all i and t. We shall be using μ as the independent variable, and with this notation (6) becomes

7) $\int_0^\infty |q_{it}(\mu)| \frac{dr}{d\mu}\,d\mu,$ using μ as independent variable, provided $r(\mu)$ is differentiable.

With these preliminaries we obtain the basic

Lemma 2: If assignment A" is a c-fold expansion of assignment A', and access perspective is homogeneous of dimensionality D, then total transportation costs incurred under A" equal $c^{(1+\frac{1}{D})}$ times total transportation costs incurred under A'.

Proof: Let T' and T" be total transport costs incurred under A' and A", respectively. From (7) we obtain

$$T' = \Sigma_i \Sigma_t \int_0^\infty |q'_{it}(\mu)| \frac{\theta}{D} \mu^{\frac{1}{D}-1} d\mu \; ;$$

$$= \Sigma_i \Sigma_t \int_0^\infty |q''_{it}(c\mu)| \frac{\theta}{Dc} \mu^{\frac{1}{D}-1} d\mu, \text{ from Lemma 1;}$$

$$= \Sigma_i \Sigma_t \int_0^\infty |q''_{it}(x)| \frac{\theta}{Dc} \frac{x^{\frac{1}{D}-1}}{c^{\frac{1}{D}-1}} \frac{dx}{c}, \text{ from the sub-}$$

stitution $c\mu = x$;

$$= c^{-(1+\frac{1}{D})} \Sigma_i \Sigma_t \int_0^\infty |q''_{it}(x)| \frac{\theta}{D} x^{\frac{1}{D}-1} dx \; ;$$

$$= c^{-(1+\frac{1}{D})} T'', \text{ from (7).} \qquad \text{QED}$$

Lemma 2 has a simple intuitive interpretation. A c-fold expansion multiplies all weights by the factor c. But it also pushes everything further out (for c>1) and so multiplies distances by the factor $c^{1/D}$. Total transport costs are then multiplied by the product of these two factors, which is $c^{(1+\frac{1}{D})}$. (The argument is not affected by c being less than one).

Up to now no element of choice or volition has been taken into account. We now assume the following situation, which is reminiscent of the first model of Section 3.5. The planner has a technology set of available land uses, any of which can be operated at any site. Net deliveries at the nucleus, q(0) are specified in advance and must be met. Subject to this restriction, the planner is to assign land uses in the Thünen pattern so as to minimize total transportation costs.

For this model we get

<u>Theorem 2</u>: If assignment A' minimizes total transport costs for the delivery schedule $q(0)$, and access perspective is homogeneous, then assignment A'', the c-fold expansion of A', minimizes transport costs for the delivery schedule $cq(0)$, for any positive c.

<u>Proof</u>: Suppose the statement is false. Then for some c, A' is optimal for $q(0)$, but A'' is not optimal for $cq(0)$. It follows from Lemma 1 that A'' is feasible for $cq(0)$. There must, therefore, be another assignment, $\tilde{A}''$, such that $\tilde{A}''$ is also feasible for $cq(0)$, and such that $\tilde{T}'' < T''$. Let $\tilde{A}'$ be the 1/c-fold expansion of $\tilde{A}''$ (i.e. $\tilde{A}''$ is the c-fold expansion of $\tilde{A}'$). From Lemma 1, $\tilde{A}'$ is feasible for schedule $q(0)$. From Lemma 2, $T'' = c^{(1+\frac{1}{b})} T'$, and $\tilde{T}'' = c^{(1+\frac{1}{b})} \tilde{T}'$. Therefore, $\tilde{T}' < T'$, which contradicts the assumed optimality of A'. QED

Note that Theorem 2 says nothing about the uniqueness of optimal land use assignments. But an easy corollary of Theorem 2 is that the number of solutions for the schedule $q(0)$ equals the number of solutions for the schedule $cq(0)$, for any positive c.

We now embed this model in a more complete one, just as was done in Section 3.5. Assume that a vector of discounted prices, $p(0)$, is given at the nucleus. This is of the same size as a, b, or q; the components $p_{it}(0)$ give the (discounted) price for commodity i at time t at the nucleus, and run over all i and all t. The prices are all per unit ideal weight for the particular i,t combinations. The inner product

$\sum_i \sum_t p_{it}(0) q_{it}(0)$--written for short as $p(0)q(0)$--is called the <u>gross value</u> of the system. It is simply the value of all inflows to the nucleus minus the value of all outflows from the nucleus, evaluated at the prescribed price vector $p(0)$. The <u>net value</u>

of the system is defined to be the gross value minus total transportation costs.

Under certain stipulations,"net value" boils down to a more familiar concept. Suppose we have a system in which identical profit-maximizing firms are bidding for land in a competitive real estate market. The profitablity of any parcel of land, under an entrepôt assumption, is given by the revenue obtained from the sale of outputs at the nucleus, minus the outlays incurred from the purchase of inputs at the nucleus, minus transportation costs both ways (all discounted to the present). This is the most that any firm would bid for the parcel, and, since there are identical firms in a competitive market, this is the amount that the parcel will actually sell for. The summation of profitability over all parcels is easily found to be nothing but "net value" itself. Under these conditions, then, "net value" is the same as total land values over the whole system. "Gross value", therefore, is the sum of total land values and total transport costs.

We assume that the planner maximizes net value. The previous criterion of minimizing transport costs subject to meeting a flow schedule at the nucleus is contained in this one, since, whatever the optimal flow $\bar{q}(0)$ turns out to be, maximizing net value clearly involves minimizing the transport costs incurred in attaining $\bar{q}(0)$. Therefore Theorem 2 may still apply, if access perspective is homogeneous.

Net value is a profit-maximization criterion, appropriate, say, to a farmer trying to extract the highest income from his land. Its relevance to a regional planner is much less clear, if it exists at all. Net value will generally be less than the present value of the corresponding income stream, since total transportation costs are subtracted to obtain net value, while ordinarily only monetary outlays, not time costs, for transportation are subtracted to obtain the income stream.

Theorem 3: Suppose that access perspective in a direct-linkage model is homogeneous of dimensionality D. The price system p(0) is given. Then any land use assignment which maximizes net value satisfies the condition: net value $= \frac{1}{D}$ total transportation costs.

Proof: Let T(q) be the function giving minimal total transport costs for all land use assignments yielding the delivery schedule $q = q(0)$ at the nucleus.

First we show that T(q) is a homogeneous function of degree $1 + \frac{1}{D}$. Suppose A' is a transport cost-minimizing land use assignment for the schedule q(0). According to Theorem 2, A", the c-fold expansion of A', is transport cost-minimizing for the schedule cq(0). According to Lemma 2, the costs for A" are $c^{(1+\frac{1}{D})}$ times the costs for A'; so a multiplication of q(0) by c multiplies transport costs by $c^{(1+\frac{1}{D})}$, and we have verified the assertion concerning T(q).

Let $\bar{q}$ be an optimal value for the delivery schedule vector q(0). The net value corresponding to this is $\bar{q}p(0) - T(\bar{q})$, and this number cannot be increased by the substitution of any other vector for $\bar{q}$. In particular, substitution of scalar multiples of $\bar{q}$ cannot increase net value. In algebraic terms, this states that $c\bar{q}p(0) - T(c\bar{q})$ must be maximized at the value $c = 1$. By the homogeneity of T, this expression equals $c\bar{q}p(0) - c^{(1+\frac{1}{D})}T(\bar{q})$.

Differentiation with respect to c yields $\bar{q}p(0) - (1+\frac{1}{D})c^{\frac{1}{D}}T(\bar{q})$. This must equal zero at $c = 1$, and so we get $\bar{q}p(0) - T(\bar{q}) = \frac{1}{D}T(\bar{q})$. The left-hand expression is net value; the right-hand expression is 1/D times total transport costs. QED

This is a remarkable theorem, since it does not depend on the technology set or on the structure of prices at the nucleus. (These prices must be taken as given, however; if the prices varied with the delivery vector q(0), the result would no longer hold).

As a special but important case, when land is controlled by identical profit-maximizing firms in an entrepôt system on a Euclidean plane, Theorem 3 states that total land values equal one-half of total transport costs.

Theorem 3, by a suitable interpretation, can be generalized to the following situation. Suppose q(0), not p(0), is given in advance, and that transport costs are minimized for this q(0). (That is, we are back in the restricted model of Theorem 2). Suppose one can find a vector, $\tilde{p}$, such that the given q(0) maximizes $\tilde{p}q - T(q)$. Then Theorem 3 tells us that $\tilde{p}q(0) - T(q(0))$ equals 1/D times total transport costs. The vector $\tilde{p}$ may be a purely artificial construct, though in some cases the "net value", $\tilde{p}q(0) - T(q(0))$, can be given a meaningful interpretation. An application of this approach will be given below.

A simple corollary of Theorem 3 is

<u>Theorem 4</u>: A structural change in a direct-linkage model which leaves access perspective homogeneous of the same dimensionality, and which results in a (rise, fall) in total transport costs, also results in a (rise, fall) in net value.

<u>Proof</u>: Net value and transport costs vary in proportion, by Theorem 3. QED

In order to find applications for Theorem 4 we must determine what kinds of structural changes leave dimensionality invariant. (In the formula $\mu(r) = (r/\theta)^D$, θ may be allowed to vary, but not D). Obvious cases are those changes not affecting either the metric or the measure--e.g. changes in the technology set or changes in prices at the nucleus. Uniform reductions in transportation cost per unit ideal weight also have this effect.

Less obvious is the following class of cases, which might be characterized by the term directionally homogeneous. Suppose one has a Euclidean metric--say a plane, for definiteness--with the ordinary areal measure, so that access-perspective is homogeneous of degree two. For each geodesic (ray) from the nucleus, let the distances be changed by a factor of proportionality; that is, all distances along the ray are multiplied by the same number. But the number itself may vary from ray to ray. The rims of the "spheres" in this new metric are homothetic images of each other (in terms of the old metric). Since these are similar to each other, area still goes up as the square of distance, and two-dimensional homogeneity is preserved.

This transformation is produced by the construction of radiating transportation arteries from the nucleus (Figure 7), provided costs are proportional to the original Euclidean distances, both on the arteries and off them. Two equi-distant rims are drawn in Figure 7, shown as homothetic,, indicating the similarity of the corresponding "spheres" about the nucleus. (Some "rough" ground is shaded in; this has the effect of drawing the rims in toward the nucleus).

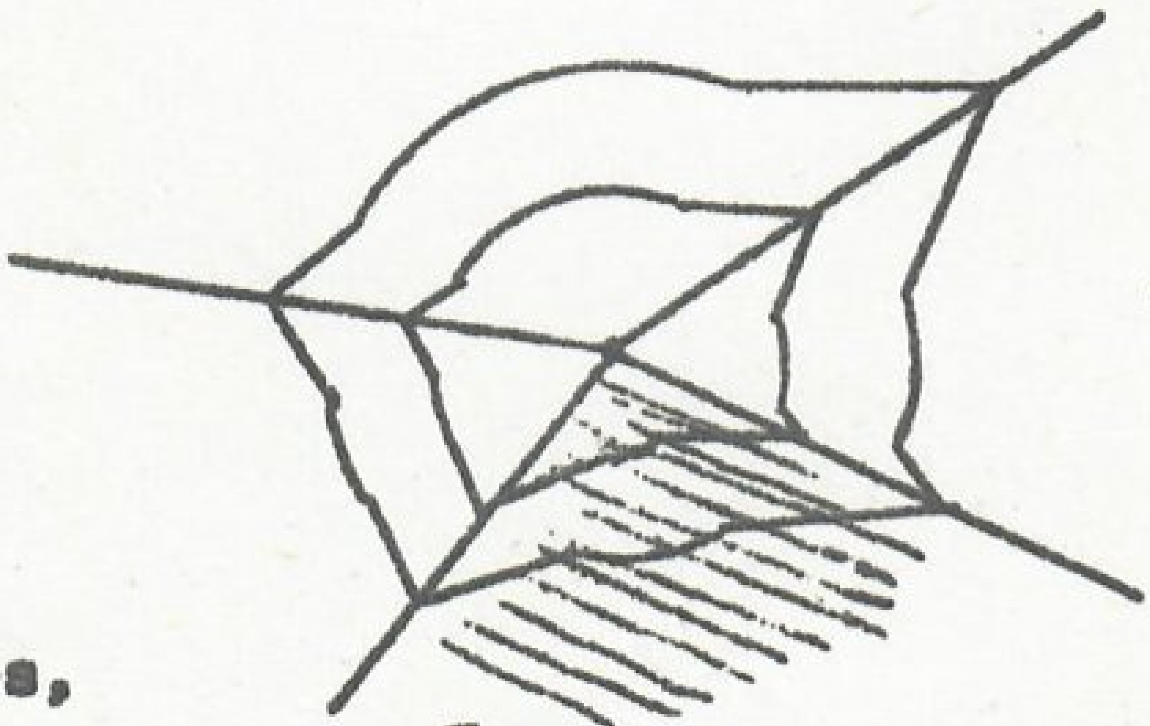

Figure 7

(One should even be able to prove that, if land is of identical "roughness" along any ray, then the access perspective will be homogeneous of dimension two--or whatever the dimension of the underlying Euclidean space. One could even let "roughness" vary with the angle of traversal--e.g. going along a road vs. crossing it--provided the cost per unit distance is independent of distance along a ray. Thus in Figure 8, movements at points a, b, and c, in the direction of the arrows make equal angles with the ray and must have equal cost per unit distance; likewise for points d, e, and f.)#

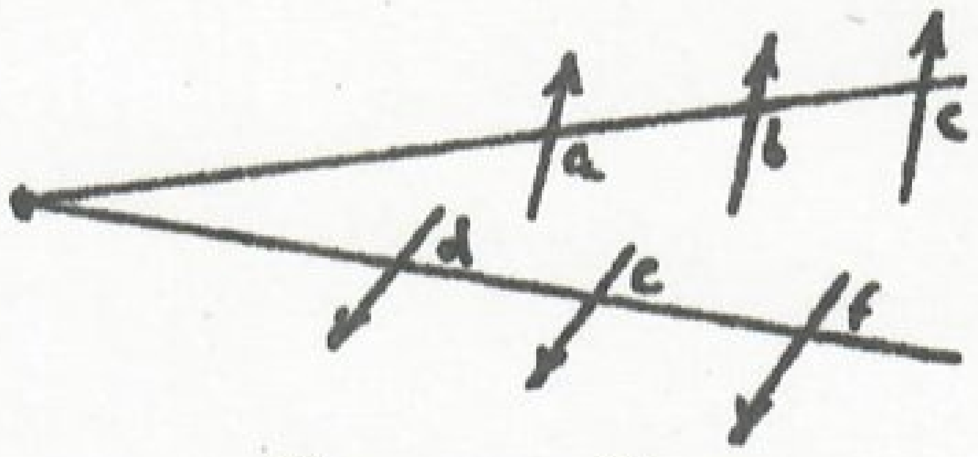

Figure 8

We shall now apply these ideas to a problem considered by Herbert Mohring.# Residents are distributed at uniform density over a circular disc about the nucleus (Central Business District), on a Euclidean plane. They commute to the CBD at a uniform rate per person. Land values (or rents, since we are dealing with a steady-state case) start at zero at the limit of settlement, and satisfy the condition that land values plus transport costs is a constant for all points of settlement. Now a transportation artery is built along a ray. Residents re-arrange themselves to

# Similar invariances under variation of cost from ray to ray have been noted by M.J. Beckmann "Bemerkungen zum Verkehrgesetz von Lardner" Weltwirtschaftliches Archiv Band 69 Heft 2, 199-213, 1952.

H. Mohring "Land Values and the Measurement of Highway Benefits" Journal of Political Economy 69:236-249 June, 1961, or H. Mohring and M. Harwitz Highway Benefits (Transportation Center, Northwestern University Press, 1962), Chapter 5.

minimize total transport costs under the new conditions, still maintaining the same uniform density and trip frequencies. The new land values and transport costs satisfy the same relations. Problem: are total land values higher before or after the transport artery has been constructed? By explicit calculation, Mohring shows that they fall as a result of the construction.

We wish to point out that this result follows immediately from Theorem 4, and that in fact a more general statement can be made. To see that Theorem 4 is applicable, one notices first that the system remains directionally homogeneous after the artery has been built, so that,before and after,access perspective is homogeneous of dimensionality two. The technology consists of just the single land use in operation, producing a single "commodity". One can find a $\tilde{p}$ such that the system behaves as if it were maximizing net value: namely, $\tilde{p}$ equals transport costs from the limit of settlement. Since total transport costs clearly decline after the artery is built, net value must decline. A moment's reflection shows that total land value, as here defined, must be the same as net value, and so it declines.

A bit more may be stated: total land value equals one-half of total transport costs, from Theorem 3. (Mohring finds this by explicit calculation). So far we have just come to the same conclusions as Mohring by a roundabout process, but now a generalization suggests itself. Any change in the transportation system which leaves access perspective homogeneous of dimension two will have the same qualitative effect as the building of a single transportation artery along a ray. For example, any finite number of radiating spokes will do, in any angular pattern. These may incorporate diverse transportation modes: a road here, a railway there, a river elsewhere. Entire sectors may be cut out, as occurs when the system fronts on a lake or ocean. The countryside may be hilly in one direction and flat in another. In all these cases, total land values will be one-half of total transport costs, so that any transportation improvement, by reducing the latter, will reduce the former. One might also generalize to the case of non-uniform densities and trip frequencies; as Mohring points out, no blanket statement can be made about the direction of the effect of transportation improvement on land values. If market demand for transportation is elastic, a transportation improvement will raise total transport costs, hence total land values.

If we confine ourselves to <u>entrepôt</u> systems, an instructive generalization of Theorem 3 can be obtained. We need a few more definitions. The contribution to net value by a unit of land at distance r from the nucleus, upon which is operating a land use of weight-density w, with input and output densities of a and b, is $\Sigma_i \Sigma_t (b_{it} - a_{it}) p_{it}(0) - rW$, which is the value of net output, evaluated at nuclear prices, minus transport costs, all per unit of area. As we mentioned above, this figure will turn out to equal land value-density in a competitive profit-maximizing market, and it is natural to abbreviate it as <u>v</u>. As a matter of fact, it can be shown that, if net value is maximized, all the theorems concerning land value-densities of the previous section remain valid. In particular, Theorem 6, which states that, as functions of r, $dv/dr = -w$, wherever w is continuous, remains valid.

We now consider a <u>truncated</u> system consisting only of a "sphere" of radius R and area M about the nucleus. This could be a complete system in which one simply runs out of land beyond distance R--e.g. an island, an impenetrable valley, or the Earth after reaching the antipodes--or it can be just a piece of a system extending beyond R to which we confine oub attention.

It will be convenient to use the area, μ, as our independent variable, rather than the distance, r. Total net value for the sphere of radius R is defined to be the contribution to net value of all the land uses located in this sphere. This equals

$$\int_0^M v(\mu)\, d\mu.$$

Total scarcity value for the sphere of radius R is defined to be Mv(M)--i.e., the area of the whole sphere multiplied by the value-density at its rim. Finally, total differential value for the sphere of radius R is defined to be total net value minus total scarcity value. (See Figure 9; the totals are the integrals $d\mu$ of the densities).#

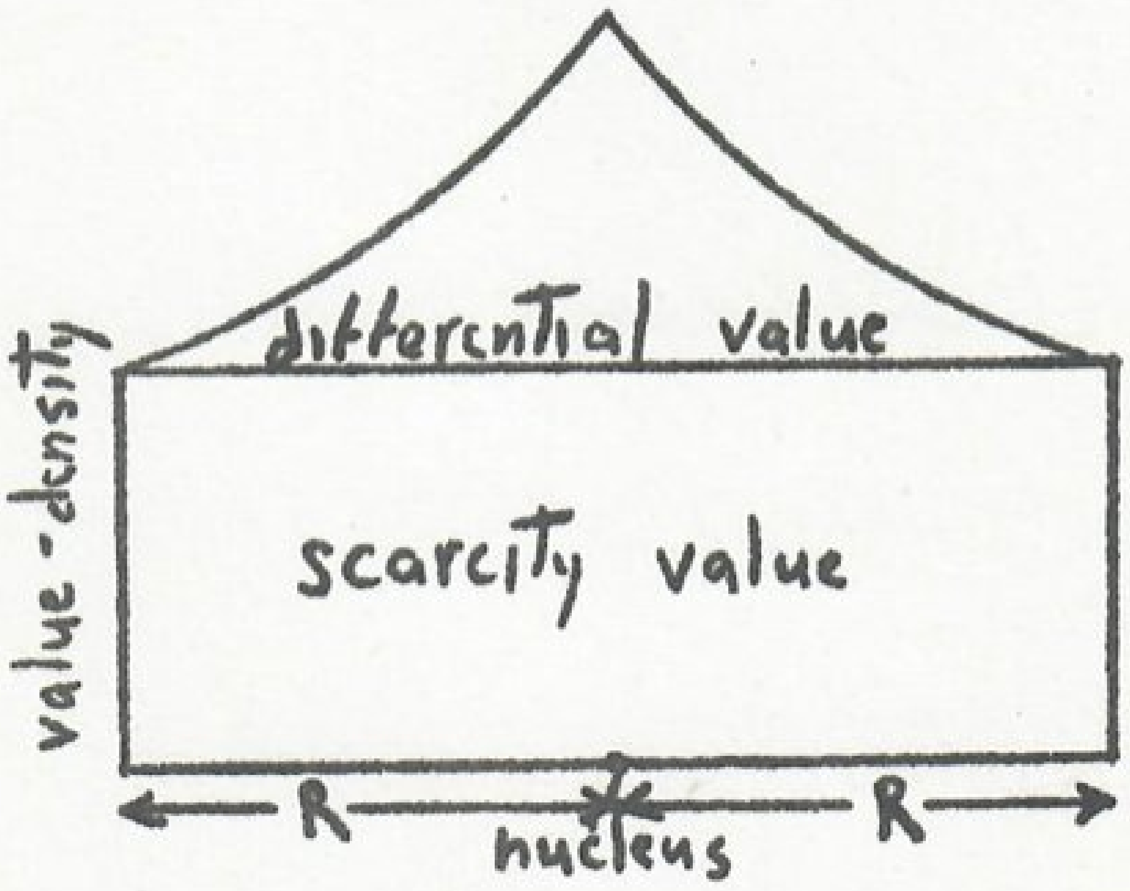

Figure 9

# For the case of the island these definitions are in accord with standard usage of the terms "scarcity" and "differential" rents, but not for the case of a truncation of a larger system. See p.85f. of B.H. Stevens "An Interregional Linear Programming Model" Journal Of Regional Science vol.1 #1, 60-98, Summer, 1958.

We shall also need the concept of total transport costs generated by the sphere of radius R, which refers only to flows between the nucleus and points in the sphere.* This equals

$$\int_0^M r(\mu)\, w(\mu)\, d\mu.$$

We are now ready to state the generalization of Theorem 3.

Theorem 5: Suppose that access perspective in an entrepôt model is homogeneous of dimensionality D, at least up to distance R. The price system p(0) is given. Then any land use assignment which maximizes net value satisfies the condition: total differential value for the sphere of radius R = 1/D times total transport costs generated by the sphere of radius R.

Proof: (This proof is given under the assumption that w(r) is continuous. The statement is true without this restriction, though a more tedious proof is then required).

Total differential value = $\int_0^M v(\mu)\, d\mu - M v(M)$.

Upon performing an integration by parts, we find that this equals $-\int_0^M \mu \frac{dv}{d\mu}\, d\mu$. By Theorem 6 of Section 4.2, $\frac{dv}{dr} = -w$. Therefore, $\frac{dv}{d\mu} = \frac{dv}{dr}\frac{dr}{d\mu} = -w\frac{dr}{d\mu}$. By the homogeneity condition, $r = \theta\mu^{\frac{1}{D}}$, so that $\frac{dr}{d\mu} = \frac{\theta}{D}\mu^{\frac{1}{D}-1}$, and

* This total may be less than transport costs incurred in the sphere, since flows from points beyond R may be passing through. We shall not use this latter concept.

$$\frac{dv}{d\mu} = -\frac{W\theta}{D}\mu^{\frac{1}{D}-1} \text{ ; so Total differential value}$$

$$= -\int_0^M \mu\left(-\frac{W\theta}{D}\mu^{\frac{1}{D}-1}\right)d\mu = \int_0^M \frac{W\theta}{D}\mu^{\frac{1}{D}}d\mu$$

$$= \frac{1}{D}\int_0^M w(\mu)\,r(\mu)\,d\mu = \frac{1}{D}\cdot \text{total transport costs.} \qquad \text{QED}$$

When $v(M) = 0$, scarcity value disappears, and Theorem 5 reduces to Theorem 3 for the entrepôt case, as it should.

So far we have been neglecting the other parameter, Θ, in the access perspective relation $\mu(r) = (r/\Theta)^D$. A moment's reflection indicates that Θ plays the role of a transportation cost parameter: a doubling of Θ doubles the cost of getting access to a sphere of a given area about the nucleus. We are now interested in discovering how the system responds to variations in Θ. The function $T(q)$, it will be recalled, gives the minimal total transport cost required to attain a delivery schedule q at the nucleus. Θ enters as a multiplicative parameter in $T(q)$. This fact could be ignored before, because Θ was held fixed, but now we write explicitly $\Theta t(q)$ for total transport costs required for q, where $t(q)$ is now free of the parameter Θ.

It is easy to see that $t(q)$ shares with $T(q)$ the fundamental property of being a homogeneous function of degree $1 + \frac{1}{D}$, if access perspective is homogeneous of dimensionality D. Our first result concerns optimal delivery schedules as a function of Θ:

<u>Theorem 6</u>: If access perspective in a direct-linkage model is homogeneous of dimensionality D, and $\bar{q}$ maximizes net value for $\theta = 1$, then $\bar{q}\theta^{-D}$ maximizes net value for any θ, prices p(0) being held fixed.

<u>Proof</u>: This follows the pattern of the proof of Theorem 2. Suppose the statement is false. Then there is a θ for which $\bar{q}\theta^{-D}$ does not maximize net value. That is, there is a $\tilde{q}$ such that

$$p(0)\tilde{q} - \theta t(\tilde{q}) > p(0)\bar{q}\theta^{-D} - \theta t(\bar{q}\theta^{-D}).$$

Multiply through by θ^{D}:

$$p(0)\tilde{q}\theta^{D} - \theta^{D+1}t(\tilde{q}) > p(0)\bar{q} - \theta^{D+1}t(\bar{q}\theta^{-D}).$$

Since t is homogeneous of degree $1 + \frac{1}{D}$, this may be re-written as

$$p(0)\tilde{q}\theta^{D} - t(\tilde{q}\theta^{D}) > p(0)\bar{q} - t(\bar{q}).$$

But this statement contradicts the assumed optimality of $\bar{q}$ for $\theta = 1$. QED

In the next few paragraphs we assume there is a unique optimal land use assignment. If this were not so, similar but slightly more complicated results could be obtained.

Theorem 6 states that, for example, in a Euclidean plane, a halving of unit transport costs would lead to a quadrupling of net deliveries at the nucleus, provided nuclear prices were not affected. With the aid of previous results we can say much more. Combining Theorem 2 with Theorem 6, we find that the response to a change in transportation parameter from 1 to θ is a θ^{-D}-fold expansion of land use assignments. With access perspective homogeneous of dimensionality D, we know that, in a c-fold expansion, all land uses spread out in such a way that distances are multiplied proportionally, by the factor $c^{1/D}$. But $(\theta^{-D})^{1/D} = 1/\theta$. We have proved .

Theorem 7: A change in transport cost parameter θ leads to an expansion of the system in such a way that corresponding distances are inversely proportional to the parameter.

Thus, a halving of unit transport costs spreads out all land uses to twice their original distances. This result does not depend on the dimensionality of access perspective. Finally, we have

Theorem 8: If access perspective is homogeneous of dimensionality D, the system elasticity of demand for transportation is $-(D+1)$. (Here,"volume of transportation" is defined as total transport costs deflated by the cost parameter θ).

Proof: Lemma 2 states that a c-fold expansion multiplies total transport costs by $c^{(1+\frac{1}{D})}$. The same applies to the volume of transportation, since θ is held fixed in Lemma 2. The induced θ^{-D}-fold expansion therefore results in a multiplication of the volume of transportation by the factor $\theta^{-D(1+\frac{1}{D})} = \theta^{-(D+1)}$. QED

This elasticity can be broken down into an elasticity of -D for the total weight to be moved, and an elasticity of -1 for the average distance to be moved. Thus in a Euclidean plane a halving of unit transport costs octuples the volume of transportation, quadrupling the tonnage and doubling the average mileage.

All these results are predicated on the assumption that the prices p(0) are unaffected by changes in the inflow-outflow schedule q(0). This assumption is appropriate for an individual farm, and also# for a small seaport which takes world prices as given.

We end this long chapter by trying to tie a few strings together. There is a marked contrast between the social equilibrium approach of Section 4.2 and the optimization approach of this section. To the extent that these approaches can be shown to be compatible, or even to imply each other, to that extent will the range of applicability of each of them be broadened. We therefore ask (1) does the entrepôt social equilibrium discussed in Section 4.2 optimize any criterion of significance? and (2) can one find a social equilibrium in a direct-linkage model which maximizes net value?

The first question is the easier to answer. According to Theorem 1 of Section 4.2, the entrepôt social equilibrium satisfies the following weight-falloff condition: If a land use of weight-density w' is operating at a point at distance r' from the nucleas, and also a use of density w" at a point at distance r", and if $r' < r"$, then $w' \geqq w"$.

The weight-falloff condition turns out to be necessarily satisfied by a solution to a certain optimization problem. To set this up we need the concept of an <u>allotment</u>. An allotment is an areal measure on the set of land uses. For example, a certain allotment might allocate two acres to turnip-growing, a hundred

# as suggested to me by William Vickrey.

acres to residences, and so forth over a class of subsets of the set of land uses. An allotment differs from a land use assignment in that the latter tells *where* each land use is to be operated, while the former merely states *how much* land is to be allocated without naming specific sites. There is a many-one relation between assignments and allotments, and a single allotment can be realized by a large number of different assignments. Two assignments corresponding to the same allotment are spatial reshufflings of each other's land uses.

Now consider the problem: Given an entrepôt system and an allotment, minimize total transport costs over the class of all assignments which realize the given allotment. Call this the *allotment-assignment problem*. We then have

Theorem 9: Any solution of the allotment-assignment problem satisfies the weight-falloff condition.

Proof: Suppose the statement is false. Then one can find a solution to the problem which violates the weight-falloff condition. That is, one can find two points, at distances r' and r'' from the nucleus, to which are assigned land uses of weight-densities w' and w'', respectively, such that $r'<r''$ and $w'<w''$. Now construct a new assignment as follows: take small spherical neighborhoods of equal area ϵ about each of these points and switch their land use assignments, leaving the assignment in all other places unchanged. This new assignment clearly realizes the allotment, since it is just a reshuffling of the solution assignment. The land uses which were switched farther away from the nucleus must ship an extra distance of $r''-r'$, and this adds an amount

$(r''-r')w'\epsilon$, neglecting higher-order terms in ϵ, to total transport costs; the land uses which were moved closer to the nucleus save an equal shipping distance, and this reduces total transport costs by an amount $(r''-r')w''\epsilon$, again neglecting higher-order terms in ϵ. The net saving is $(r''-r')(w''-w')\epsilon$ (plus higher-order terms), and for small enough ϵ this must be positive. But this contradicts the assumption that the original assignment minimized total transport costs. QED

The converseof Theorem 9 is also true: any assignment realizing the allotment and satisfying the weight-falloff condition solves the allotment-assignment problem.

As for the second question: a social equilibrium maximizing net value can easily be found in the special case of entrepôt systems. Identical profit-maximizing entrepreneurs who bear transport costs between their sites and the nucleus will maximize net value, given prices at the nucleus and their common technology set. This result follows from the two facts (1) net value is the integral of net profit per unit area, and (2) the profitability at a given point is completely independent of what goes on elsewhere in the system.

Much more difficult is the problem in the direct-linkage case, because the independence mentioned in point (2) breaks down. The following discussion is meant to be suggestive rather than rigorous. It points to the conclusion that profit-maximizing entrepreneuws, with identical technologies, acting under competitive conditions, will maximize net value in the general direct-linkage case, just as they do in the entrepôt case.

We assume that a system of local prices is generated. $p_{it}(r)$ is the (discounted) price of commodity i at time t at distance r from the nucleus. For short, the whole system of prices at r will be written p(r). The nuclear prices p(0) are simply the local prices at the nucleus. p(0) is special in that it is given, while all other price structures are built up by the action of agents in the system. The controller of each site selects that land use which maximizes his profits, the latter being calculated on the basis of the local prices at that site. That is to say, the controller of a site at distance r solves the problem

$\underset{A}{\text{maximize}} \; \Sigma_i \Sigma_t \, (b_{it} - a_{it}) \, p_{it}(r)$, where A ranges over his technology set, and the b's and a's are the output and input densities for land use A. All individuals have the same technology set.

Finally, prices and flows are assumed to satisfy the efficiency conditions of Section 2.4. For the direct-linkage situation this means (1) $\left|\frac{\Delta p_{it}}{\Delta r}\right| \leqq 1$; (2) if $q_{it}(r) > 0$, then $\frac{d p_{it}(r)}{dr} = -1$; (3) if $q_{it}(r) < 0$, then $\frac{d p_{it}(r)}{dr} = 1$. Condition (2) states that, if there is an inflow toward the nucleus, price declines with distance; condition (3) states the corresponding fact for outflows.

These last two paragraphs state conditions that might reasonably hold in a regime of profit-maximizers under direct-linkage conditions. We will now show that these conditions also arise if an overall planner were to maximize net value. If these conditions are sufficient to determine a solution maximizing net value, we would be assured that the competitive social equilibrium did in fact maximize net value. To derive these conditions, which are necessary rather than sufficient for a solution, we shall use the Maximum Principle of Pontryagin.*

Omitting details, we may outline the principle as follows. One is to maximize $\int_{r_1}^{r_2} f(q(r), A(r))\,dr$, where r is a real variable, q and A are vector functions of r, f is a real-valued function; the integral limits r_1 and r_2 may also be subject to choice, but for our purposes they may be taken as fixed. The functions q(r) and A(r) satisfy the system of differential equations $dq/dr = g(q(r), A(r))$. The element of choice enters in the function A(r), which may be any bounded measurable function whose range lies in a certain closed set, τ, which is specified in advance. This choice, together with initial conditions on q, determines the entire course of q(r), hence f and the integral.

This sets up the problem. The Maximum Principle introduces an auxiliary vector function, $\psi(r)$, and forms the Hamiltonian function $H = f + \psi g$. (The last term is an inner product). The following relations are then satisfied by the optimal solution. (1) For every r, the optimal choice of A is one maximizing H over the possible choices $A \varepsilon \tau$. (2) $\psi(r)$ satisfies the system of differential equations $-\frac{d\psi}{dr} = \frac{\partial H}{\partial q}$.

* L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze, and E.F. Mischenko The Mathematical Theory of Optimal Processes (K.N. Trirogoff, translater; New York, Interscience, 1962).

We now translate the problem of maximizing net value into the Pontryagin form. r, q, and A will retain their customary meanings of distance from nucleus, traffic flow, and land use, respectively. The set $\mathcal{T}$ from which A is to be drawn is simply the technology set. The differential equations dq/dr=g(q(r), A(r)) are, when written out in full, $\frac{dq_{it}}{dr} = -(b_{it} - a_{it})\frac{d\mu}{dr}$,

where b_{it} and a_{it} are output and input densities for the land use A in operation at distance r. This follows immediately from Theorem 1 of this section. Finally, the integral to be maximized is

$$\int_0^{\infty} \Sigma_i \Sigma_t \left[p_{it}(0)\left(b_{it}(r) - a_{it}(r)\right)\frac{d\mu}{dr} - |q_{it}(r)| \right] dr.$$

This expression perhaps needs some explanation! Comparison with Theorem 1 shows that the integral of the first term is nothing but $\Sigma_i\Sigma_t p_{it}(0)q_{it}(0)$, which is gross value. Comparison with formula (6) shows that the integral of the second term is minus total transportation costs, so the entire expression does indeed equal net value.

We now form the Hamiltonian:

$$H = \Sigma_i \Sigma_t \left[p_{it}(0) - \psi_{it}(r)\right]\left[b_{it}(r) - a_{it}(r)\right]\frac{d\mu}{dr} - \Sigma_i \Sigma_t |q_{it}(r)|.$$

Let us abbreviate $p_{it}(0) - \psi_{it}(r)$ as $p_{it}(r)$, for r>0. It can be shown that $\psi_{it}(0) = 0$ for all i and t, so this notation extends to r=0 as well.

The first part of the Maximum Principle states that the optimal land use at any distance is one maximizing H, or, equivalently, maximizing $\Sigma_i \Sigma_t p_{it}(r)(b_{it}(r) - a_{it}(r))$ over the technology set. But this is exactly what oub profit-maximizing controllers at distance r will do, __if__ the $p_{it}(r)$ may be identified with the local prices which they face. To justify this identification it ought to be shown that these $p_{it}(r)$ satisfy the efficiency conditions (1), (2) and (3).

The second part of the Maximum Principle states that

$$-\frac{d\psi_{it}}{dr} = \frac{\partial H}{\partial q_{it}}. \quad \text{But } \frac{\partial H}{\partial q_{it}} = -\frac{d|q_{it}|}{dq_{it}}$$

$= -1$, if $q_{it} > 0$; $= 1$, if $q_{it} < 0$; indeterminate, if $q_{it} = 0$. This gives us

$$\frac{dp_{it}(r)}{dr} = -\frac{d\psi_{it}}{\partial r} = \begin{cases} -1, & \text{if } q_{it} > 0 \\ +1, & \text{if } q_{it} < 0 \\ \text{indeterminate}, & \text{if } q_{it} = 0. \end{cases}$$

The first two parts of this triple statement are precisely the efficiency conditions (2) and (3). By considering smooth approximations to the absolute value function, it might also be shown that condition (1) holds as well.

This concludes our discussion. It has been shown that a certain social equilibrium system imitates the optimality conditions for maximizing net value in a direct-linkage model, as given by the Maximum Principle. (This discussion is interesting in another way. It shows that a principle designed for optimization over time can also be used for optimization over space).

5. Selected Short Subjects

5.1. Building Height and Real Estate Values

We present here a simple model relating the intensity of land use, the value of unimproved lots, and the value of improvements. "Intensity" is phrased in terms of building heights, but the model is perhaps applicable to such quantities as residential population density or intensity of fertilizer use.

Suppose that a syndicate acquires control of a vacant lot, upon which it erects an N-story building. The size of the building, and the use to which it is put, are chosen to maximize present value. For simplicity, it is assumed that N can assume any non-negative value. If, in fact, N can assume only integer values, the error committed will be at most half a story.

The present value of an N-story building, together with its lot, when put to its most profitable use, and assuming no further construction is undertaken on the lot, is assumed to be

$$1) \qquad N\left(h - \frac{k}{2}N\right) = R(N),$$

where h and k are parameters depending on the location of the lot. In this model, k is assumed to be fixed for all sites, but h--which may be called the <u>rentability</u> of the site--is assumed to vary, being generally higher on sites closer to the center of town. The cost of construction of an N-story building is assumed to be

$$2) \qquad a + bN + \frac{c}{2}N^2 = C(N), \text{ for } N > 0; \; C(0) = 0;$$

where a, b, and c, are again parameters which may depend on location, but in this model are assumed to be the same for all sites.

Formula (2) arises from the following construction. There is a set-up cost of <u>a</u> for initiating any improvement. The marginal cost of the N-th story is $b + cN$ (or $b - c/2 + cN$, if N assumes only discrete values). That is, part of the cost is the same for each additional story, and part is proportional to its height above the ground (due to vertical transportation costs, elevator shafting, etc.) Formula (1) is a quadratic approximation, with the additional specification that no improvements at all bring in no net revenue.

Of the five parameters--a, b, c, h, and k--the first four will, in general, be positive. The sign of k is less certain. On the one hand, the costs of vertical transportation diminish the value of lofty heights, and tend to make k positive; on the other hand, higher stories escape the noise and fumes, and offer a commanding view "far from the madding crowd"--considerations which tend to make k negative. (It would appear, in general, that marginal value at first rises with height, but then declines for very high stories. To incorporate this would require the use of at least a cubic in N, and so it was avoided). These opposed tendencies suggest that it would be worthwhile to explore the special case k =].

We will assume that a, b, and c, are positive, and also that $c + k > 0$, and that $h - b > \sqrt{2a(c+k)}$. This last relation turns out to be the condition that optimal N be positive, and the next-to-last relation is the condition that optimal N be finite.

If construction time is negligibly short--as we will assume--then the value of the vacant lot upon which an N-story building is erected, is simply $R(N) - C(N) = V(N)$. In a competitive market of identical profit-maximizers, the selling price of the lot will be the maximal attainable value of V(N). Assume there are no zoning restrictions or other constraints on N. Then the selling price of the lot will be $\max_{0 \leq N < \infty} V(N) = V(\overline{N}) \quad \overline{V}$.

We are interested mainly in the ratios of R, C, and V, to each other, and how these respond to changes in the parameter of rentability, h. There are, in fact, some interesting empirical regularities here. The ratio V/C tends to rise as one approaches the center of town.# Thus for residential land uses the value of the unimproved lot is generally a small fraction of the cost of the building placed on it, while for the highest-priced uses in the heart of the central business district, a rule-of-thumb is that the value of the lot and the cost of the building placed on it should be about equal.#

It seems reasonable that central locations have high rentabilities compared with peripheral locations.

Theorem 1: If N is adjusted to maximize profits, $\bar{V}/\bar{C}$ rises with h.

F.C.R. Douglas Land Value Rating (London, Hogarth Press, 1936), p.2f.

S.L.McMichael and R.F. Bingham City Growth and Values (Cleveland, Stanley McMichael Publishing Organization, 1923), p. 115, 136.

Proof: Maximizing V over N, we find that $\bar{N} = \frac{h-b}{c+k}$.

By substitution, $\bar{V} = \frac{(h-b)^2}{2(c+k)} - a$, and

$\bar{C} = \frac{c(h-b)^2}{2(c+k)^2} + \frac{b(h-b)}{c+k} + a$. Therefore,

$$3) \qquad \frac{\bar{V}}{\bar{C}} = \frac{\frac{1}{2(c+k)} - \frac{a}{(h-b)^2}}{\frac{c}{2(c+k)^2} + \frac{b}{(h-b)(c+k)} + \frac{a}{(h-b)^2}} .$$

Since $h > b$, a rise in h increases the numerator of (3) and decreases the denominator. QED

Theorem 2: If $k = 0$, and N is adjusted to maximize profits, then, as h goes to infinity, the ratio $\bar{V}/\bar{C}$ approaches 1.

Proof: In formula (3), the limit approached as $h \to \infty$ is obviously

$$\frac{\frac{1}{2(c+k)}}{\frac{c}{2(c+k)^2}} = \frac{c+k}{c}. \quad \text{For } k=0, \text{ this } =1. \qquad \text{QED}$$

The conclusion of Theorem 1 conforms with the evidence./ The conclusion of Theorem 2 is the rule-of-thumb $\bar{V} = \bar{C}$ for very high values of h.

A further interesting result of this model is

Theorem 3: As h goes to infinity, the number of stories goes up asymptotically as the square-root of land value.

Proof: $\bar{N}$ is a linear function of h, while $\bar{V}$ is a quadratic functtion of h. QED

5.2. Some Problems of Intra-Urban Location

Three general problem types will be considered in this section. (1) Within the general confines of an entrepôt system, where will land uses which are linked directly to the outside world locate? (2) Suppose a headquarter point is to service an entire Thünen system; where will it locate? (3) What can be said about the shape and location of neighborhoods within a city?

We start with an entrepôt system, so that all land uses are linked exclusively to the nucleus. A new land use is introduced which trades directly with the outside world in addition to its trade with the nucleus. (An example would be a manufacturing plant whose exports do not all go through the nucleus). It is assumed that this new land use occupies a negligible quantity of space in relation to the whole system, so that the overall pattern of land uses and land values is not disrupted.

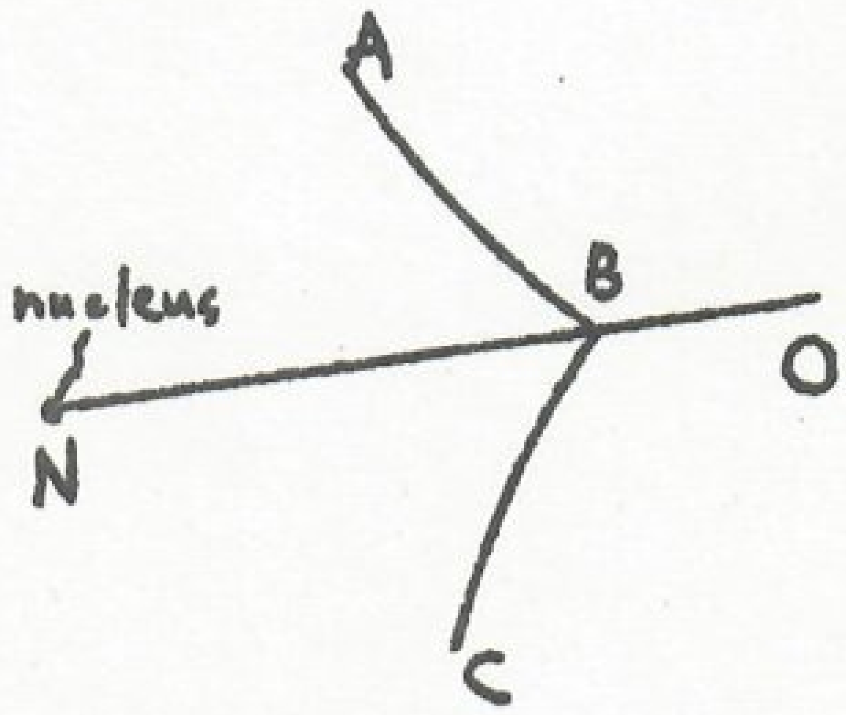

Figure 1

Now suppose there is a system of transportation arteries radiating from the nucleus, as in Figure 7 of Section 4.3 or Figure 4 of Section 2.5. One of these--NBO--is shown in Figure 1. The line ABC crossing the artery has all its points equi-distant from the nucleus.

Just as any other land use, our new one obeys the site-substitution principle in locating; that is, it minimizes the sum of transport costs and land values. There are two components to transport costs here: costs incurred for flows to and from the nucleus, and costs incurred for flows to and from the outside world. Thus there are three components to consider in the site-substitution criterion.

Let us now compare the three points A, B, and C with each other. Transport costs to the nucleus will be the same for all three points, since they are equi-distant. Land values will be the same at all three points, according to Theorem 3 of Section 4.2. The only remaining distinction which can arise is transport costs incurred on trade with the outside world. Under the great majority of plausible circumstances, point B will be closer to the outside world than points A or C, provided these are close enough to B. (For example, this will occur when the geodesics from A and C to the outside world have stretches in common with the artery NBO, and the underlying metric is Euclidean).

It follows that points such as A and C are eliminated as potential sites for our new land use, since they are dominated by point B. We conclude that land uses having direct links with the outside world, when situated in an entrepôt system having radiating transport arteries, will locate somewhere <u>on</u> the arteries.

In this argument we have been assuming that the new land use occupies a negligible amount of land. Suppose this were not so. It is easy to see in a qualitative way what will happen. Competition among outside-linked land uses for land along the transport arteries will drive up land values there compared to land values at non-highway-oriented locations. The pure entrepôt-linked land uses will be driven away from points such as B.

The general pattern that emerges is that land uses which trade exclusively with the nucleus will locate off the arteries, at points such as A and C. To these the conclusions of Section 4.2 still apply: the weight-falloff condition and convex decreasing land value-densities. Strips of land along the arteries, having higher value-densities than points at equal ideal distances off the highways, will by occupied by land-uses linked in part to the outside world.

For example, manufacturing plants which import or export materials from outside the local community will locate on river fronts, railways or highways. Tourist-catering retail trade, entertainment places, hotels, etc., are strung out along major radial highways.

We now come to the second problem. A certain city is a Thünen system, so that points equi-distant from the nucleus carry the same land uses. Into this symmetrical arrangement is to be placed a headquarter point serving the entire system. (Examples would be the city hall, the central post office, or the municipal airport, provided there is to be just one city hall, post office, or airport for the entire city). The land to be occupied by the headquarter point is compact and small enough to be thought of as a geometrical point, but--unlike the situation of Chapter 3--not so small that land values can be neglected.

The level of services provided by the headquarter point is to be uniform in the following sense. All points carrying the same land uses receive the same level of services. Actually, for our purposes, we may make the even weaker assumption: all points equidistant from the nucleus receive the same level of services (per unit area). That is, there is a function $g(r)$ giving the level of services per unit area at distance r from the nucleus.

We shall assume that the metric is rectangular, as discussed at the end of Section 3.6. Without loss of generality, the nucleus may be placed at the origin of a pair of co-ordinate axes.

By convention, let us assume that the X-axis runs from West to East, and the Y-axis runs from South to North (See Figure 2). The point with co-ordinates (x,y) is at distance $|x| + |y|$ from the nucleus, and the"spheres" about the nucleus are the 45°-tilted square discs whose diagonals are coincident with segments of the axes.#

To clarify a possible ambiguity in Figure 2, it should be noted that the axes do <u>not</u> represent low cost transport arteries. If they did, we would obtain a more complicated <u>non</u>-rectangular metric. The spheres for this latter metric are depicted in Alonso <u>Location and Land Use</u>, op. cit, p.132.

Let the headquarter point be located at $(\bar{x},\bar{y})$. We assume that headquarters is directly-linked to all other sites in the system, so that the distance between it and the point with co-ordinates (x,y) is $|x - \bar{x}| + |y - \bar{y}|$. (Whether the other points are directly-linked to each other or not is irrelevant for our purposes).

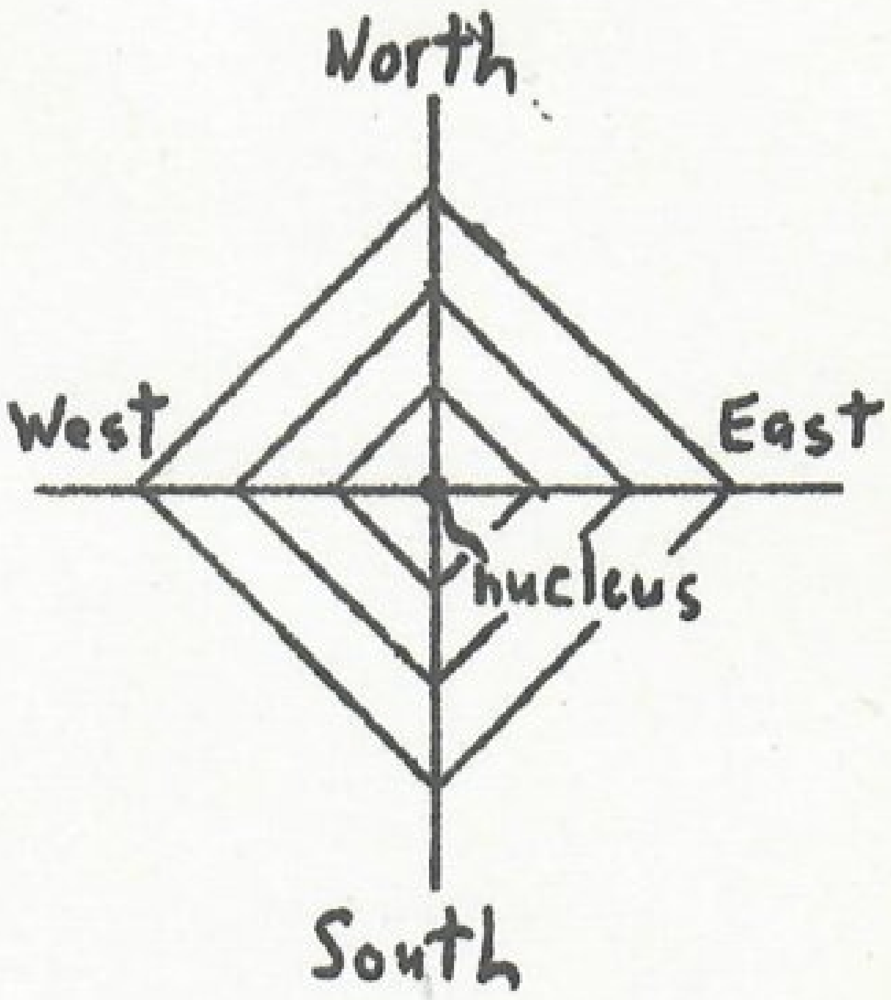

Figure 2

We now invoke the site-substitution principle. The headquarter point is to be located so as to minimize the sum of total transport costs plus land values.

Total transport costs is a fairly complicated expression in $\bar{x}$ and $\bar{y}$. However, for our purposes we need only the fact that it is a convex function of position: for $0 < \lambda < 1$,

$$T(\lambda\bar{x}_1 + (1-\lambda)\bar{x}_2, \lambda\bar{y}_1 + (1-\lambda)\bar{y}_2) \leqq \lambda T(\bar{x}_1, \bar{y}_1) + (1-\lambda) T(\bar{x}_2, \bar{y}_2).$$

Lemma: For any non-negative measure over the plane (representing required delivery levels) total transport costs under a rectangular metric are a convex function of the position of the headquarter point.

Proof: Suppose the measure is positive at just a single point, (x', y'). Total transport costs are proportional to $|\bar{x}-x'| + |\bar{y}-y'|$, which is easily seen to be a convex function in $(\bar{x}, \bar{y})$. Now suppose we have an arbitrary discrete distribution. Total transport costs for this is the sum of transport costs for each of the points of the distribution. But a sum of convex functions is a convex function, and the result is established for any discrete distribution. Finally, let us take a continuous distribution (or a mixed discrete-continuous distributicn). Approximate this by a discrete distribution in the following way. Partition the plane into regions such that no region has a diameter larger than some number ϵ. Place anywhere in each region a single point having a mass equal to the measure of the original distribution in the region. It is easily established that, for any headquarter location, the error made in estimating total transport costs for the original distribution by using the discrete approximation is of the order of ϵ. Now take a sequence of approximations with $\epsilon \to 0$. For each of the approximations, transport costs are a convex function of position. But the limit of a sequence of convex functions is a convex function, which establishes the result in general.* QED

The same argument, word for word,--with the exception that $\sqrt{(\bar{x}-x')^2+(\bar{y}-y')^2}$ is substituted for $|\bar{x}-x'|+|\bar{y}-y'|$ --establishes the result for the Euclidean metric, a fact which was stated without proof in Section 3.3. Cf. H.W. Kuhn and R.E. Kuenne "An Efficient Algorithm for the Numerical Solution of the Generalized Weber Problem in Spatial Economics" Journal of Regional Science vol. 4, #2, 21-33, Winter, 1962, especially p.25f.

The main result may now be derived.

Theorem: An optimal headquarter location exists on the 45°-slope diagonals running from South-west to North-east and South-east to North-west, if it exists at all. (These are dashed in Figure 3).

Proof: Consider the points A, B, C, D, and E, which lie symmetrically placed about the SE-NW diagonal in Figure 3. These points are all equidistant from the nucleus, and so they all have the same land value-density. Therefore, only transport costs influence their relative desirability as headquarter locations. By symmetry of the whole distribution, transport costs at A and E are equal, transport costs at B and D are equal, and the same applies to any pair of points equidistant from C along this line. Suppose B were optimal; then D would also be optimal, by symmetry; but then C, on the diagonal,

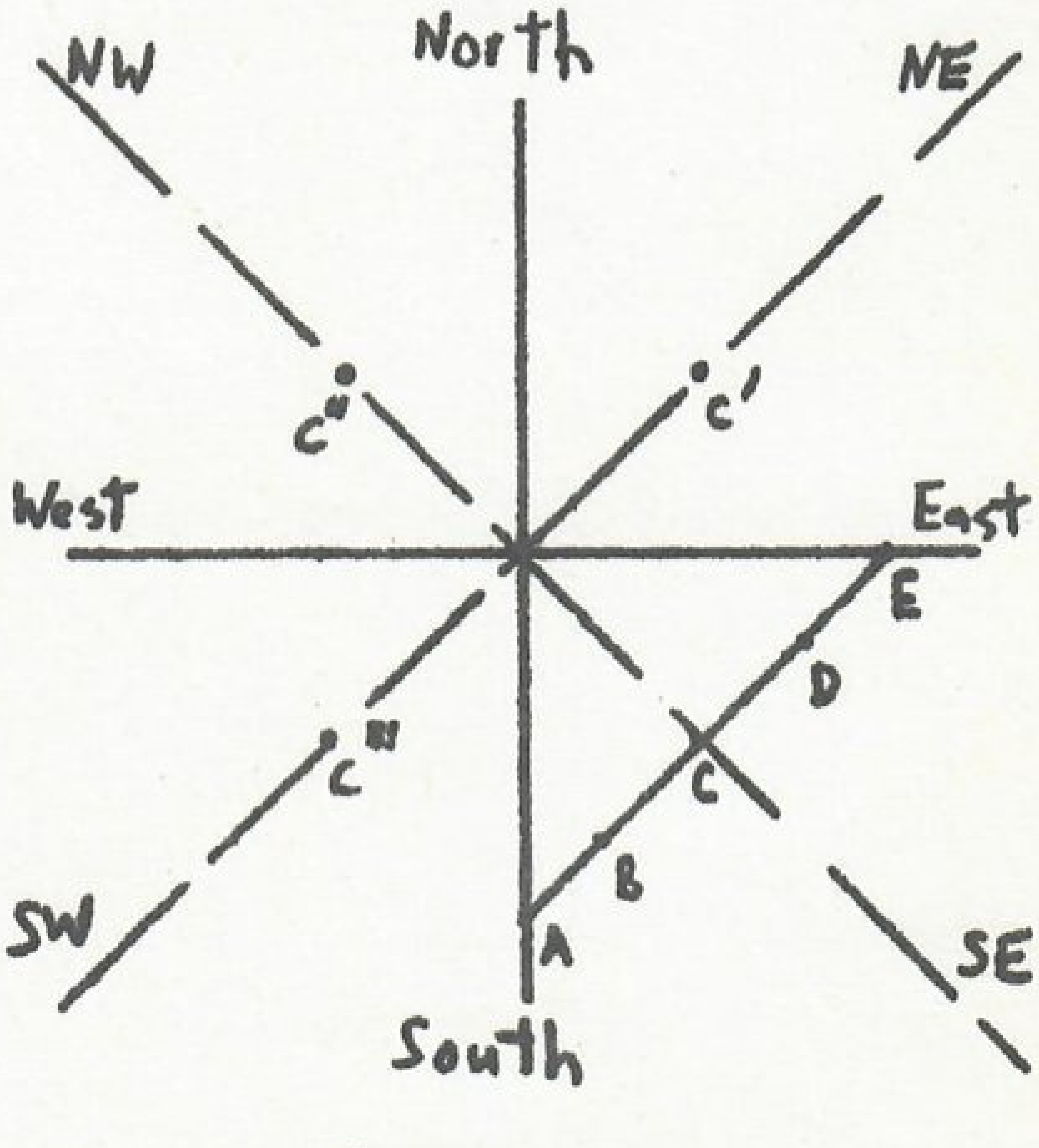

Figure 3

would be optimal, too, by the convexity of the total transport cost function. We have proved that, if any location were optimal, then its perpendicular projection on the nearest diagonal would also be optimal. QED

This theorem does not prove that an optimal cannot be off the diagonal as well. To establish the stronger result that only the diagonal contain optimal points, one must establish the strict convexity of the transport cost function, at least in the vicinity of the diagonals. It may be shown that this local strict convexity holds at the points $(\pm \bar{x}, \pm \bar{x})$ if $\int_{\bar{x}}^{\infty} g(r)\,dr > 0$.

By symmetry, there are always at least four optimal headquarter locations. For example, if point C is optimal in Figure 3, the points C', C", and C"' are also optimal.

Nothing has been said, so far, about how far out along the diagonal the optimal location should lie. A standard comparative-statics argument shows that, the denser the headquarter land use is, the closer to the nucleus it will lie--a not unexpected result in view of the weight-falloff condition for entrepôt systems. Thus, a light land use such as an airport will lie far out along a diagonal, while a dense use such as a central post office will lie close in.

We now come to the third problem: neighborhood shapes and locations. This is a very intricate problem, and the discussion which follows is more in the nature of a study of approaches and lines of attack than a presentation of results.

A <u>neighborhood</u> is a fairly coherent region of similar land uses. This definition is rather vague, in that degree of "similarity" required is not specified, nor degree of "coherence". If the same type of land use is found on both banks of a river, or both sides of a street, for example, we may for some purposes think of the river or street as demarcating the boundary between two neighborhoods, and for other purposes ignore them and use a more inclusive neighborhood concept.

In Thünen systems the neighborhoods are rings, or parts of rings, enclosing the nucleus. This formation arises, in entrepôt systems at least, from the operation of two forces on the individual land users: the attraction of the nucleus, and the repulsion from regions of high land value. We now wish to superimpose upon these forces other kinds of forces, which will have the effect of distorting or disrupting the ring formation caused by the first two acting alone.

One broad category of additional forces is made up of restrictions on land uses or land users. Let us take zoning laws, for example. If the chosen land use on site L in the absence of zoning resprietions would be A, and A is not prohibited by the law, then A will still be chosen; the restriction is not binding. If A were prohibited, the alternative permitted use making the highest bid for the site will be the one chosen, and land value will, in general, be lower than it would be in the absence of zoning.*

* This statement may appear incompatible with the often-advanced thesis that zoning preserves or enhances land values. If neighborhood effects had been included in our analysis, this could indeed happen. The prohibition of a land use on one parcel could enhance the value of adjacent parcels, and since this effect is

reciprocal, it is possible for all values to rise, even though the direct effect of restrictions on land use in a parcel is to depress its value. For more on the effects of zoning see Alonso, op. cit., Chapter 6.D.

In general, the effects of excluding a land use type from a certain territory are--in addition to driving it from that territory--to expand the amount of remaining area devoted to that use, and to shift the distribution of uses within the type toward more intensive uses; (e.g., exclusion of a minority group from one part of a city increases crowding and leads to "invasions" of other parts of the city). The entire process may be called pseudomorphosis of the land use type, on the geological analogy.

A second broad category is made up of forces of attraction of similar land uses upon each other.* There is a variety of underlying causes producing such attraction. The first, and simplest, is the "birds-of-a-feather" phenomenon, the desire of people to associate with their own kind. A second cause is informational economy. For example, a foreign-language enclave may form, not necessarily because people have any direct preference for the company of their co-linguists, but simply because the ease of communication makes it that much easier to get on. Also, communication-oriented industries, which depend heavily on the flow of

* This and the final category to be discussed lead to voluntary segregation of land uses, as opposed to the perhaps involuntary segregation induced by restrictions. Cf. G.S. Becker The Economies of Discrimination, op. cit., p.59.

information among their component firms, tend to cluster.*

* See E.M. Hoover and R. Vernon Anatomy of a Metropolis (Garden City, New York, Doubleday Anchor Books, 1962), p.59ff.

A third cause is the economy of pooling. Dense concentrations of complementary resources smooth out demand fluctuations for each of them.*

* P.S. Florence The Logic of Industrial Organization (London, Kegan, Paul, 1933), Chapter 1; Hoover and Vernon, loc. cit., p.62.

How is this raw material to be worked up into a model? One might try to derive the relations for individual optimization for each of the land users of the neighborhood, and solve them simultaneously. These would be highly interdependent, owing to the attraction of land uses upon each other. A possible short-cut is the following approach. We assume the entire neighborhood locates so as to optimize a criterion. For example, suppose a land use allotment is given (See Section 4.3) which the neighborhood is to realize. Subject to the constraint of realizing the allotment, maximize: "self-attraction of the neighborhood" minus total transport costs incurred by the neighborhood minus total value of the land occupied by the neighborhood (if devoted to alternative uses). "Self-attraction of the neighborhood" is given by

$$4) \quad \iint f(r(x,y))\, d\mu(x)\, d\mu(y),$$

where μ is, say, the population measure of the inhabitants of the neighborhood, x and y range independently over the points of the system, r is distance, and f(r) is a decreasing real-valued function. f(r) represents the "utility" contributed by a pair of people at distance r apart.

This is certainly not an easy problem to solve, but one may make plausible conjectures as to what the solution looks like in certain simple cases. In particular, let us take the entrepôt case on the Euclidean plane. We would conjecture that the solution to the stated problem has the general form of the shaded area of Figure 4: somewhat elliptoid in boundary, topologically a disc, symmetric about a radial spoke from the nucleus, and elongated in a direction perpendicular to this ray. The intuitive argument for this conjecture is the following. If only transport costs and land values had received consideration, the neighborhood would be a circular Thünen ring about the nucleus. If only the "self-attraction of the neighborhood" received consideration, the neighborhood would be (presumably) a circular disc. The neighborhood of Figure 4 is a "compromise" between these two extremes. It should also lie roughly at the distance from the nucleus that the Thünen ring would have been.

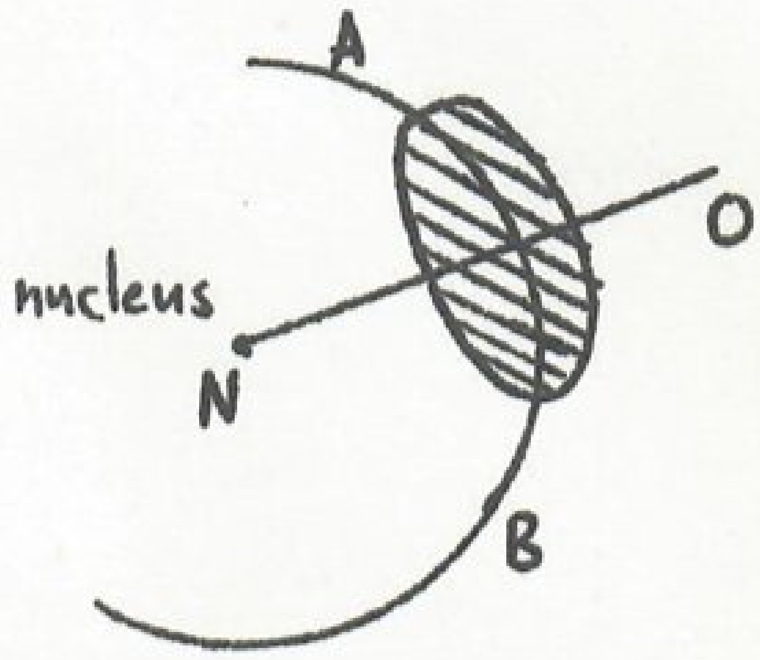

Figure 4

The pattern of land values will also be distorted from the decreasing convex function of distance in the entrepôt case. If the attraction of neighborhood people on each other is strong enough, there should emerge a local peak of land value-density somewhere in the middle of the neighborhood, since this will presumably be the "coziest" part of the neighborhood, and competition for the middle part will drive up land values there.

A rather different category of forces are those which form neighborhoods, not by attracting uses or users to each other, but to a third object which is attractive to all activities of a certain type. The Thünen rings themselves may be thought of as formed by a force of this type, the common attraction being the nucleus.

Points of access to the transportation system, such as local railway stations, serve as nuclei for the formation of small Thünen systems within the overall context of larger ones. Points of access to major utilities, such as water or power sources, may play a similar role. But the neighborhoods thus formed are rather heterogeneous, since access to transportation or utilities is an attractive force on all land uses. To explain the formation of neighborhoods of a fairly homogeneous character, we need objects which attract only a limited range of land uses.

Consider a facility, located at a site L, whose services appeal only to people in a limited taste range. These people tend to move toward site L, and the resulting rise in land values about L (and fall in land values in the places these people vacate) tends to drive away people who are indifferent to the facility at L. The argument can be extended from this dichotomous case to the case where there is a distribution of tastes over the population. Measure the net attractiveness of the site to a person by the marginal rate of substitution of land value-density for distance from site L.* There will be a tendency for people to ar-

* A full-blown location theory for entrepôt systems has been constructed by William Alonso from the indifference maps of "bid-price" curves from which these marginal rates of substitution

are derived. See Alonso, op. cit., Chapters 3-5.

range themselves so that the most attracted are the closest, and the least attracted (or most repelled, perhaps) are the farthest, from site L.

For example, stores specializing in luxury goods will attract the rich. Churches of different denominations scattered about give rise to a tendency for clusterings to occur by their respective parishoners around each, and for non-churchgoers to move away from them all. In general, a process of voluntary self-segregation of people by taste and income classes tends to occur.

Not only does the existence of diverse facilities lead to a process of segregation, but conversely, segregation leads to the existence of facilities. A facility will not be constructed without there being a certain threshold demand level concentrated at a point, and this will be passed when a sufficient number of people with a taste for the services of the facility are concentrated at sufficient density in the vicinity of the point. The cumulative effect of this interaction may be much greater than that of the first process acting alone, in creating diverse neighborhoods.*

* Cf. W.R. Thompson A Preface to Urban Economics, op. cit., p.128; C.M. Tiebout "A Pure Theory of Local Expenditures" Journal of Political Economy 64:416-424, October, 1956. In Tiebout's analysis, the self-segregated neighborhoods that are formed are local political jurisdictions, and the corresponding diverse facilities which serve them are the bundles of public goods offered by each.

Any significant scale economies in the production of services above the point of threshold demand accentuates the effects of this interaction.

In attempting to set up at least a partial model for these processes, we might start with (4), the expression for self-attraction of a neighborhood. Since people are not attracted to each other, but to the facility located at site L, (4) might be replaced by the simpler expression

5) $\int f(r(x, L))\, d\mu(x)$. *

Given an allotment, one might now maximize the criterion: expression (5), minus transport costs, minus land values, and arrive at a solution not too dissimilar to Figure 4. Site L may be fixed in advance, or it, too, may be subject to choice. When the scale and type of services to be provided by the facility at L are also variable, this relatively simple approach appears to break down, and a re-formulation along the lines of Section 3.5 may be in order. Not only are scale and spacing of service facilities to be chosen, but also type of services, and the distribution of population over space by tastes. Work on models of this

* Alonso has used (5), with μ uniform, as the expression for the utility of an individual lot to its user, where the integral is taken over the lot area, and L is the location of the user's "front door". See Alonso, op. cit., Appendix B. Without discussing the merits of this approach, we suggest it is at least as applicable to the problem of neighborhood location as it is to individuals.

complexity has scarcely begun.

Alonso has pointed out that self-attractive neighborhoods will tend to string themselves out along transport routes, because of the high internal accessibility which such a route affords. If both radial routes (such as NO in Figure 4) and crosstown routes (such as AB in Figure 4) are available for such purposes, it is plausible that the crosstown route would be preferred. Elongation along a crosstown route allows the neighborhood to remain close to its "natural" distance from the nucleus, as determined by the weight-falloff condition. But elongation along a radial route would place a good portion of the neighborhood too close, and a good portion too far, from the nucleus. This observation weakens the case for Alonso's equilibrium explanation of the "sector theory" of urban growth.# The latter specifies radially elongated neighborhoods.

If the land allotment is not given in advance, some interesting interactions between self-attraction, density, and radial distance of neighborhoods from the nucleus occur. Let us take as an example that very special neighborhood--the office-building complex of the central business district. This neighborhood has a strong self-attraction, due to the face-to-face contact requirements for negotiations, etc. This fact, coupled with the easy stackability of administrative activities, leads, as we know, to the use of very dense, high-rise office buildings, and this in

Alonso, ibid., p.140-142. The "sector theory" was proposed by Homer Hoyt in The Structure and Growth of Residential Neighborhoods in American Cities (Federal Housing Administration, 1939).

turn leads to a very centralized location for the whole complex in an overall entrepôt system, by the weight-falloff condition. An innovation which loosens these bonds of attraction (e.g., high-fidelity closed-circuit television) would reduce the density at which office activities were carried on, and decentralize the neighborhood.

5.3. Police-Criminal-Victim Equilibrium

We have said almost nothing so far about the spatial patterns that result from the interaction of several population groups. Here we develop a simple model involving three groups: a population of potential Victims, a population of potential Criminals, who commit crimes upon the Victims when the opportunity presents itself, and a population of Policemen, who try to prevent Criminals from perpetrating their misdeeds. The type of crime which this model fits best is robbery, though the basic elements are present in a whole spectrum of crimes.

The entire region of study is assumed to be divided into n precincts, the area of the i-th precinct being M_i, the area of the whole being $M = \sum_i M_i$. In the i-th precinct there are V_i victims, C_i criminals, and P_i policemen; there are V, C, and P in the whole region, respectively, so that $V = \sum_i V_i$, $C = \sum_i C_i$, $P = \sum_i P_i$. All three groups may move among the precincts, but the totals are fixed. The population density of victims, criminals, and police in the i-th precinct is $v_i = V_i/M_i$, $c_i = C_i/M_i$, and $p_i = P_i/M_i$, respectively. (All of the M_i and V and C are assumed to be positive; some of the other quantities may be zero). For simplicity, we assume that these quantities can take on any real values, rather than being confined to integers. Let K_i be the crime rate in the i-th precinct (i.e. total crimes per unit time), $k_i = K_i/M_i$ the crime rate density, and $K = \Sigma_i K_i$ the regional crime rate.

We want an expression for the crime rate, K_i, in the ith precinct, in terms of the other quantities. It is assumed that our three sub-populations wander independently and at random within the i-th precinct. Let us first take the case where there are no policemen present. We assume that a crime occurs whenever a chance "encounter" takes place between a criminal and a victim. The simplest reasonable assumption is that the expected number of encounters per unit time per unit area is proportional to the density of victims, and to the density of criminals.* By choosing units appropriately, we can make the constant of proportionality equal 1, so that we get

6) $K_i = c_i v_i M_i = c_i V_i = C_i v_i = C_i V_i/M_i.$

Now let us introduce policemen. Some of the chance encounters will not lead to crimes if there is a policeman nearby to inhibit the criminal. Again making the simplest assumption, we postulate that an encounter will lead to a crime if and only if no policeman is present within a certain distance of the point of encounter. The probability of this absence decreases exponentially with the density of policemen. By choosing the unit of area appropriately, we can make this factor equal e^{-P_i}, where e is the base of natural logarithms. By assumption, this event is indepen-

* Cf. the Law of Mass Action in chemistry, in which the reaction rate of two substances is proportional to the product of their molecular densities.

dent of the occurrence of the encounter, and so the expression in (6) must be multiplied by this factor to get the expected crime rate. Thus we get

7) $K_i = c_i v_i M_i e^{-P_i}$

as the complete expression for the crime rate in precinct i.

These crime rates motivate the distribution of our three populations. For victims, the most relevant data are crime rates per victim for each precinct, since this gives the expected rate at which crimes will be committed against any individual person. This is $K_i/V_i = c_i e^{-P_i}$, and victims will tend to move from precincts where this datum is high to precincts where it is low.

For criminals, the most relevant data are crime rates per criminal for each precinct, since this determines expected "income" of criminals in that precinct. This is $K_i/C_i = v_i e^{-P_i}$, and criminals will tend to move from precincts where this datum is low to precincts where it is high.

As for the police, we assume they distribute themselves so as to minimize the total crime rate K, taking into account possible repercussions of their moves on the distribution of victims and criminals.

Let us first consider the no-police case. We assume that victims move between precincts whenever there is a chance to reduce the crime rate per victim exposure by so moving. A criminal will move to another precinct whenever the other one has a higher crime rate per criminal than the one he is in now. We then arrive at the following simple result.

Theorem 1: If victims move to precincts with lower crime rates per victim, and criminals move to precincts with higher crime rates per criminal, and there are no policemen, then the only equilibrium solution is the one where the population density of victims is the same for all precincts, and the same is true for criminals.

Proof: Crime rate per victim in the i-th precinct equals c_i, and crime rate per criminal equals v_i. If the c_i's are all equal, no victim has an incentive to move; if the v_i's are all equal, no criminal has an incentive--so this at least is an equilibrium solution. To see that it is the only solution, suppose that another equilibrium solution existed with $c_1 > c_2$. There cannot be any victims in precinct 1, else they would leave; but if there are no victims the criminals will leave, which contradicts our assumption. Also, suppose another solution existed with $v_1 > v_2$. There cannot be any criminals in precinct 2 in this case; but then all victims would go to precinct 2, giving another contradiction. QED

This could easily be turned into an explicit dynamic model --for example, by making migration rates for victims proportional to the origin-destination differential in crime rates per victim, and similarly for criminal migration. It then turns out that the equilibrium solution of Theorem 1 is stable--in fact, globally stable. Intuitively this may be seen as follows. The precincts of greatest victim density will be receiving a criminal influx, which causes victims to leave eventually. The precincts of lowest criminal density will be receiving a victim influx, which causes criminals to enter eventually.

The simple equilibrium of Theorem 1 has a rather intriguing property. Suppose two people, Vic and Crim, are playing a zero-sum game, in which the pay-off to Crim is the total crime rate, K. Vic has a total quantity V which he can deploy in any way among the n precincts, whose areas M_i are given. Crim has a total quantity C with the same freedom, and the pay-off for each precinct is given by (6). Then the equal density solution for each gives the unique saddle-point among pure strategies.

To show this, we note that

$$\min_{\{V_i\}} K = \min_{\{V_i\}} \Sigma_i \, c_i V_i = \left(\min_i c_i\right) V \leqq \frac{CV}{M}, \quad \text{and}$$

$$\max_{\{C_i\}} K = \max_{\{C_i\}} \Sigma_i \, C_i v_i = C\left(\max_i v_i\right) \geqq \frac{CV}{M}.$$

The first inequality becomes an equality if and only if all the densities c_i are equal; the second inequality becomes an equality if and only if all the densities v_i are equal. This establishes the result.

More interesting results are obtained when policemen are admitted. Criminal movements are as above, toward precincts of high crime rates per criminal. We shall now assume, however, that the distribution of victims is fixed, unaffected by the actions of criminals or police. That is to say, the incidence of crime is assumed to be of negligible importance in determining the distribution of the general population of victims, compared with other forces. We again have a two-person game of sorts, this time between criminals and police, rather than criminals and victims.

Formally, two people, Pol and Crim, are playing a zero-sum game, in which the pay-off to Crim is the total crime rate, K, which equals $\sum_i v_i C_i e^{-P_i/M_i}$, from formula (7). Crim can choose any non-negative quantities C_i, subject to the constraint $\sum_i C_i = C$, where C is given; Pol can choose any non-negative quantities P_i, subject to the constraint $\sum_i P_i = P$, where P is given. For the police, this game-theoretic behavior is assumed explicitly, since they are to be deployed to minimize the total crime rate. For criminals, it must be verified separately that their independent actions lead to the same result as would occur if some underworld mastermind were deploying them.

It turns out that a far-reaching characterization of the solution to this game can be given. To arrive at this we need the following preliminary result.

<u>Lemma</u>: Let $P>0$, $M_i > 0$, $v_i > 0$, for $i = 1, 2, \ldots n$. Consider the following two minimization problems; in both cases the minimization is to be over non-negative quantities P_i, subject to the constraint $\sum_i P_i = P$. (1) Minimize: $\text{Max}_i\ v_i e^{-P_i/M_i}$; (2) Minimize: $\sum_i v_i M_i e^{-P_i/M_i}$. Both these problems have the same unique solution.

<u>Proof</u>: Let λ be the minimal attainable value for the objective function of problem (1). Then $\lambda \geqq v_i e^{-\bar{P}_i/M_i}$, for all i, and, if $\bar{P}_i > 0$, then $\lambda = v_i e^{-\bar{P}_i/M_i}$. The first relation is obvious. To prove the second, suppose it were false for some i'; for this i', we would have $\lambda > v_i e^{-\bar{P}_i/M_{i'}}$; if $\bar{P}_{i'}$ were reduced slightly, and the other $\bar{P}_i$'s increased in compensation, λ would be reduced, contradicting its assumed optimality. Given P, these relations uniquely determine λ and the $\bar{P}_i$.

Turning to the second problem, we note that the derivative of the objective function with respect to P_i is $-v_i e^{-P_i/M_i}$. According to the Kuhn-Tucker Theorem,* there is a number, which may also be called λ, such that $\lambda - v_i e^{-\bar{P}_i/M_i} \geqq 0$, for all i, and, if $\bar{P}_i > 0$, then $\lambda - v_i e^{-\bar{P}_i/M_i} = 0$. But these are the same relations as in problem (1). Since, as we have noted, these have a unique solution, the lemma is established. QED

With the aid of this result, the existence and uniqueness of a solution to the police-criminal game may be established in Theorem 2. The detailed nature of this solution will be spelled out in Theorem 3.

<u>Theorem 2</u>: The zero-sum game with pay-off to Crim of $\sum_i v_i C_i e^{-P_i/M_i}$, as described above, has a pure-strategy saddle-point. The strategy for Pol is unique, and, if, for all i, $v_i \neq \lambda$ (where $\lambda = \underset{\{P_i\}}{\text{Min}} \left[\text{Max}_i \; v_i e^{-P_i/M_i}\right]$), then the optimal strategy for Crim is also unique.

<u>Proof</u>: The existence of the saddle-point will be established by proving the following chain of equalities and inequalities.

$$C\lambda = \underset{\{P_i\}}{\text{Min}} \; \underset{\{C_i\}}{\text{Max}} \; K \geqq \underset{\{C_i\}}{\text{Max}} \; \underset{\{P_i\}}{\text{Min}} \; K \geqq \underset{\{P_i\}}{\text{Min}} \sum_i v_i \bar{C}_i e^{-P_i/M_i}$$

$= \in\lambda$, where $\bar{C}_i$, i=1, 2,..n, is a particular strategy for Crim which will be specified below. The truth of these relations

* H.W. Kuhn and A.W. Tucker "Non-linear Programming", p.481-492 of <u>Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability</u> (Berkeley, University of California Press, 1951).

implies that the weak inequalities are in fact equalities, and this shows that the pair $\{\bar{C}_i\}\{\bar{P}_i\}$ is a saddle-point, where $\{\bar{P}_i\}$ is any minimizer of $\Sigma_i v_i \bar{c}_i e^{-P_i/M_i}$.

Of these relations, the two inequalities stem from general properties of the "Max" and "Min" operators. The first equality is established as follows. Suppose that a deployment of policemen is given. All criminals will move to those precincts for which crime rate per criminal, $v_i e^{-P_i/M_i}$, is a maximum. This will yield a total crime rate of $C \operatorname{Max}_i v_i e^{-P_i/M_i}$. The minimization of this over police strategies is $C \min_{\{P_i\}} \left[\operatorname{Max}_i v_i e^{-P_i/M_i}\right] = C\lambda$, which is the first equality.

To establish the last equality, we choose the following strategy for Crim. In the problem--Minimize: $\operatorname{Max}_i v_i e^{-P_i/M_i}$, there is a unique solution, according to the Lemma. Let I be the set of precincts for which $P_i > 0$ in this solution. All criminals are to be placed in I, in such a way that the density of criminals is the same for all these precincts. That is, for $i \in I$, $c_i = C/M'$, where M' is the total area of the precincts in I, and $c_i = 0$ for $i \notin I$. Then $\Sigma_i v_i \bar{C}_i e^{-P_i/M_i} = \frac{C}{M'} \Sigma_{i\in I} v_i M_i e^{-P_i/M_i}$. This is to be minimized over $\{P_i\}$. By the Lemma, this has the same solution as the problem -- Minimize : $\operatorname{Max}_{i\in I} v_i e^{-P_i/M_i}$. For this problem, $\bar{P}_i > 0$ for $i \in I$, and so $v_i e^{-\bar{P}_i/M_i} = \lambda$. (The ristriction of i to I makes no difference). Therefore, $\frac{C}{M'} \Sigma_{i\in I} v_i M_i e^{-\bar{P}_i/M_i} = \frac{C}{M'} \Sigma_{i\in I} M_i \lambda = \frac{C}{M'} M' \lambda = C\lambda$, which is the equality.

We have thus established the fact that the following pair of strategies is a saddle-point: for Pol, the deployment that solves the problems of the Lemma; for Crim, an equal density at all precincts for which there are police, and no criminals elsewhere.

The uniqueness of the Pol strategy follows from the uniqueness of the solution to problem (1) of the Lemma--any deviation will raise the value of $C \operatorname{Max}_i v_i e^{-\bar{P}i/Mi}$.

Finally, let us consider alternative optimal strategies for Crim. There are two types possible. The first is a redistribution of criminals among the precincts of I, so that criminal densities are no longer equal. It is not hard to show that a shift of police from low to high criminal density precincts will reduce the total crime rate below $C\lambda$, so this is non-optimal for Crim. The second type is a shift of criminals outside of I. If, for all $i \notin I$, $v_i < \lambda$, such a shift automatically reduces the crime rate. In this case, the saddle-point strategy for Crim is, therefore, uniquely optimal. QED

(In the knife-edge case where $v_i = \lambda$ for some i, it can be shown that there are, indeed, multiple solutions for Crim. This case seems hardly worth exploring. In all other cases, police and criminals are either both present in a precinct, or both absent.)

We now characterize this saddle-point solution in detail. For simplicity, it will be assumed that the knife-edge case mentioned above does not occur, so that everything is uniquely determined. Let λ be as above, the minimal value of the objective function of problem (1) of the Lemma.

<u>Theorem 3</u>: In the game-theoretic equilibrium, precincts fall into two radically different regimes, depending on whether their victim densities, v_i, are less than, or greater than, λ. All precincts for which $v_i < \lambda$ have no police, no criminals, and no crime, and nothing more need be said about them. The rest of this theorem refers to the precincts for which $v_i > \lambda$. Police density is given by: $\bar{p}_i = \log(v_i/\lambda)$. The following quantities are the same for all these precincts: density of criminals, crimes per unit area, and crimes per criminal (the latter being equal to λ). Crimes per victim are <u>inversely</u> proportional to the density of victims.

<u>Proof</u>: If a precinct has police, the equation $\lambda = v_i e^{-\bar{p}_i}$ must be satisfied (see the proof of the Lemma). This is impossible if $v_i < \lambda$, so none of these precincts have police. In the saddle-point, there are criminals only where there are police, so criminals are absent, too. This takes care of the first type of precinct. Precincts for which $v_i > \lambda$ must have police, for otherwise the relation $\lambda \geqq v_i e^{-\bar{p}_i}$ would be violated. Since they have police, $\lambda = v_i e^{-\bar{p}_i}$, and so $\bar{p}_i = \log(v_i/\lambda)$. Crimes per criminal in the i-th precinct $= v_i \bar{C}_i e^{-\bar{p}_i}/\bar{C}_i = v_i e^{-\bar{p}_i} = \lambda$, a constant. The constant density of criminals is given by the saddle-point. These last two facts imply the constancy of crimes per unit area. This last fact implies that crimes per victim are inversely proportional to the density of victims. QED

The results of Theorem 3 are presented graphically in Figure 5. The independent variable is density of victims for a precinct, increasing from left to right. The vertical scales of the five functions depicted are chosen for convenience, and are not comparable with each other. All five functions are zero to the left of $v=\lambda$, so that four of them are discontinuous at λ. Apart from this drastic change of regime, the most surprising result is the curve for crimes per victim. The victims who are worst off are the ones living at medium densities, just above λ.

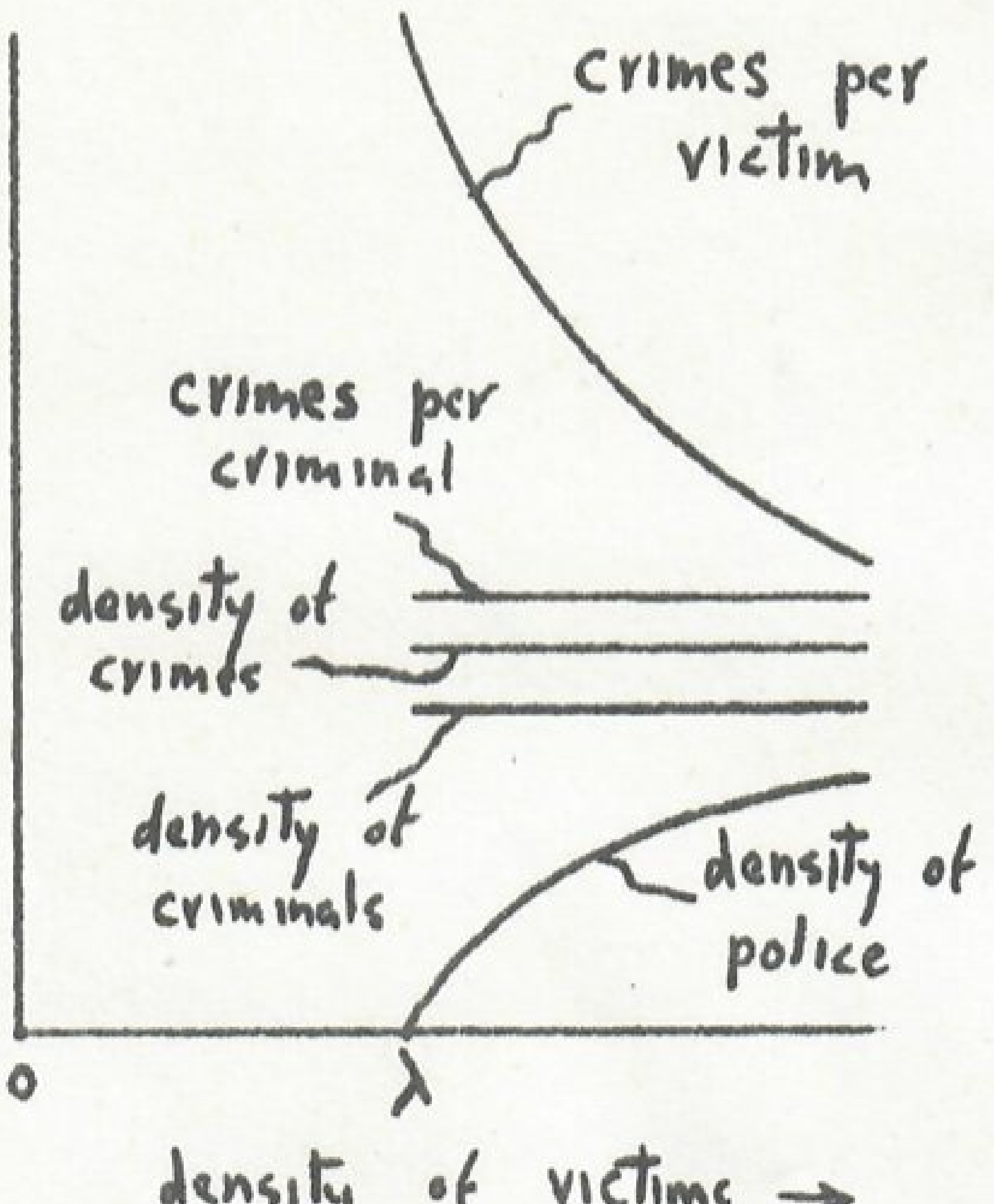

Figure 5

We may now investigate the effects of changes in the parameters of the model--the v_i's, M_i's, C, and P. Just a few cases will be considered here. It is easily verified that a uniform doubling of victim population doubles λ, doubles crime rates, and has no effect on the deployment of police or criminals by precinct. A doubling of the total criminal population, C, merely doubles crime and criminals everywhere, and has no effect on the deployment of policemen.

More interesting is the effect of a rise in the total number of policemen, P. Of course this leads to a fall in the total crime rate, but it does so by reducing λ. An unfortunate precinct whose victim density is just below the old λ level, and just above the new one, will find that crimes per victim jump from zero to the highest level in the region--a result of increased law enforcement! The explanation, of course, is that criminals spread into "greener pastures" when police become too numerous in the more densely populated precincts.

Law enforcement is often alleged to have beneficial "spillover" effects--for example, through the apprehension of criminals who might prey on other communities. This effect certainly exists, but the present model--which does not deal with the apprehension of criminals--points up the existence of a spillover effect in the opposite direction, which may be more important. Stricter law enforcement induces potential criminals to emigrate to other communities.

A community whose population density ($\doteq$ victim density) rises may find itself in the midst of a crime wave when the critical density λ is crossed. Many suburban communities seem to be in this situation.

INDEX