BARUCH SCHOOL C.C.N.Y

NOTES ON A THEORY OF POPULATION DISTRIBUTION

The spatial distribution of population is closely linked with the distribution of geographical features, natural resources, mining, agriculture, manufacturing, transport and trade, so that when one tries to explain the former one is led to the more comprehensive task of explaining the distribution of economic activity in general. The most striking feature of all these distributions is their extreme unevenness their tendency to agglomerate into isolated homesteads, then hamlets, towns, cities, metropolitan areas, and even more comprehensive concentrations. The two explanatory approaches are by means of (a) location theory, which is a branch of economic analysis, and applies the usual maximization principles to location decisions; (b) stochastic processes, which generate observed distributions by postulating (simple) laws of redistribution over time. The two should be taken as complementary and not mutually exclusive.

Location theory

We may say, in very general terms, that observed distributions result from the tendency for complementary factors and vertically linked activities to get close to each other; (and--much less important--for activities that interfere with each other to stay out of each other's way). The measure of the attractive force between two activities is the saving in transfer cost which would result from their closer proximity. Transfer costs may be broken down as follows: (a) costs to the thing transferred--opportunity costs, quality deterioration, risk of accident; (b) costs to the transferring structures--construction, maintenance and direct operating costs. The structures include storage facilities, transmitters, channels, receivers; the channels may or may not have moving parts. The kinds of circulation which occur in a modern economy are really quite varied. A partial list follows:

thing circulated	storage	transmitter	stationary channel	moving channel	receiver
freight, people	warehouse station	--	tracks, roadbed	trains	--
" "	"	--	highway	cars trucks	--
water	reservoir	pumps	conduits	--	
electricity	--	dynamo, transformer	wires	--	motor, light-bulb
people, mail	airport		air	airplane	airport
gas, oil	tanks	pumps	pipelines	--	--
information	brains, files	telephone	wires	--	telephone
people	house, job	--	sidewalk	--	--
garbage	dump	pumps	sewers	--	ocean, chemical plant
freight, people	harbor		water	ship	harbor
" "	street level		shaft	elevator	floor level
cattle	range	--	trail	cowboys	stockyards

The list, as stated, is partial. It also overlaps--for example, the circulation of people is also to some extent the circulation of the information they possess. The transfer of printed matter involves both freight and information, and goes through several successive kinds of circulating media.

The task of assessing overall transfer costs is rather formidable: (a) firstly, because of the great variety of circulating media, and of component costs of each;

(b) secondly, because transfer activities are the realm par excellence of natural monopolies, public regulation and subsidy, and direct government operation; we may therefore expect that prices charged to the buyers of transfer services need bear no close relation to costs incurred by the sellers; that the cost of durable investments need not be close to the revenue they yield; and also that the aim of profit maximization be less compelling (to the sellers) than in other parts of the economy. We may mention the following instances: the system of tolls and freeways on public highways, railroad price discrimination against goods of high value per unit bulk, irregular zone tariffs in railroads. Other practices reflect a partial, but probably poor, response to true cost differentials: relatively low rates on long-hauls, bulk shipments, full-carload lots, staple commodities, to or from big centers, on low-demand return hauls; night and Sunday telephone rates; higher rates per ton on bulky, perishable, noxious, dangerous or fragile commodities; and many other examples.

We are thus caught in a dilemma. On the one hand location theory is based on the existence of transfer costs, and its applicability to the real world would seem to rest on an approximately correct appraisal of these costs; on the other hand these costs are quite hard to come by. There are several different approaches we may use to rescue ourselves.

(a) We may, in spite of all, use a simple cost function--the simplest being that costs are constant per ton-mile. We then deduce a slew of results and compare with the real world. If the conformity is good, we take this as evidence that our simplifications were not too drastic.

(b) We may use a function flexible enough to be realistic, and yet not unmanageable. For example, I have worked with the function Cost $= \alpha + \beta M + \gamma T + \delta MT$ where M is miles and T is tons. This may be thought of as the beginning of a Taylor series in M and T; it has four degrees of freedom and yet implies some strong results. A function of the form Cost $= \alpha + T(\beta + \gamma M - \delta M^2)$ may fit the facts even better, though I have not worked with it as yet.

(c) We may use a simple function and then ask how the results would be perturbed if we substituted what we think is a more realistic one. Frequently, the qualitative form of the result would not change; only some of the numerical relations would change.

(d) Finally, it turns out that many results would hold under a very wide variety of cost functions. For these there is no problem.

Land-Use Models

In industrial location theory, it is customary to start with an activity and ask where it will be located. In agricultural location theory it is customary to start with a location and ask what it will be used for. For a number of reasons the second approach is preferable even for the industry problem. We accordingly start with a model which has the happy properties of being simple in conception, fruitful in results, wide-ranging in applicability, and realistically valid to a high degree. It is also the oldest in location theory, dating from 1826 (JH von Thunen, Der Isolierte Staat). Imagine a single city located on a uniform plain. Uniformity means that there are no inhomogeneities of climate, soil, topography, mineral deposits or man-made structures. In this economy, isolated from the world, there are a number of activities present in the technology of the people. Let a_{ij} represent the number of tons of commodity i required as input for unit level of activity j ; let b_{ij} be the number of tons of output of commodity i from unit level of activity j; let p_i be the price of commodity i in the city; let t be the cost of transportation per ton-mile, assumed uniform for all commodities, weights and distances. All activities are to take place

on the land surrounding the city, which itself is a point. All inputs are to be transported from the city to where the activity is carried on, and all outputs are to be transported from the activity site to the city.

Two measuring conventions will simplify the analysis. It is convenient to define the unit level of an activity as that requiring one unit of land, say one acre. To run activity j at level x then requires xa_{ij} and xb_{ij} tons of input and output of commodity i and also x acres of land. Does this preclude the possibility of varying intensity of land use? No, since different intensities are conceived to be different activities.

The second convention may be illustrated by an example. Suppose it costs 2 ¢ to transport a ton of coal a mile, and 6 ¢ to transport a ton of cotton (because of its bulkiness). We may then say that one ton of cotton has an ideal weight of three tons, using coal as the standard; that is, as far as transport costs are concerned, one ton of cotton is equivalent to three tons of coal. (This usage goes back to Alfred Weber, 1909.) Again, suppose it costs 4¢ a mile to transport a commuter (because of his demand for space and comfort), and that his opportunity cost is $1 per hour and the train averages 40 mph; complete costs to him are then 6½ ¢ per mile, which gives him an ideal weight of 3¼ tons. One may assign an ideal weight to information units (letters or bits) by the same procedure. Since only ideal weights are relevant for location theory, actual weights are ignored: all input and output coefficients are assumed measured in ideal weights. (One application: suppose the transport cost of a commodity is halved--say for oil by pipelines; this is equivalent for location purposes to halving the oil input and output coefficients of all activities into which it enters, and doubling its price.)

A particular acre of land will be used for that activity which pays it the greatest rent, provided this be positive; otherwise it stands unused. The surplus which an activity has to pay rent is given by the excess of the value of its outputs over the value of its inputs. The farther from town it is, the less valuable its outputs and the more valuable its inputs will be, and so the lower its surplus will be.

Let us define for an activity j the two quantities V_j and C_j as follows:

$$V_j = \Sigma_i p_i(b_{ij} - a_{ij}), \qquad C_j = \Sigma_i (a_{ij} + b_{ij}).$$

V_j is what the surplus from activity j would be if it were carried on in the city, i. e., with city prices applying to inputs and outputs. C_j is the total (ideal) weight of inputs plus outputs; it is the total weight carried to and from the city and so measures the rate at which total transport costs rise with increasing distance from the city. It is contended that the two numbers V_j and C_j sum up everything of locational significance in this simple model for activity j. The rent-paying capacity of activity j at distance d from the city is $V_j - tdC_j$. The activity which will be carried on at that distance is the one which maximizes the foregoing expression. Since only distance, and not direction, are involved, it follows that activities will array themselves in concentric circular bands centering on the city.

Consider the borderline distance between two consecutive activities. At the border, the rent-paying capacities of the two activities must be equal. If we move further outward, both rent-paying capacities decline, but the one with the smaller C_j declines by less, so that it can outbid the other; conversely, if we move inward, it increases by less, and so is outbid by the other. It follows that activities are ordered from innermost to outermost by decreasing C_j. If we relabel activities so that the innermost is 1, the next 2, and so on, this may be written $C_j > C_{j+1}$.

We may next find the distance from the city of the borderline between two successive activities by equating their rents V - tdC and solving for d. This yields

$$d_{j,j+1} = \frac{1}{t}\,\frac{V_j - V_{j+1}}{C_j - C_{j+1}}$$

Since this must be positive, it follows that $V_j > V_{j+1}$ — that is, V_j also must decrease as we move outward. The rent at the borderline may be found by substituting the obtained value for d into $V_j - tdC_j$ which yields

$$\text{rent} = (C_j V_{j+1} - V_j C_{j+1})/(C_j - C_{j+1}).$$

Since this must be positive, it follows that $(V_{j+1}/V_j) > (C_{j+1}/C_j)$—that is, the C's must decrease at a faster relative rate than the V's do as we move outward. Finally, if we take three successive activities, the C's and V's must be such that the distance from the city of the borderline between the second and third exceeds the distance of the borderline between the first and second; also, the rent on the latter must exceed the rent on the former. These two requirements both lead to the same constraint, namely $C_{j-1}V_j + C_j V_{j+1} + C_{j+1}V_{j-1} > V_{j-1}C_j + V_j C_{j+1} + V_{j+1}C_{j-1}$.

Deeper investigation shows that total rent equals one-half of total transport costs.

Suppose that transport costs per ton-mile, t, fall, with no change in the V's (infinitely elastic demands and supplies). No change in the kinds of activities employed or their ordering will occur. The distance formula shows that all borderlines will move outward in proportion to the % fall in t. The area devoted to each activity will increase as the square of this proportion. It may be shown that the elasticity of demand for transportation is -3, comprised of a tonnage demand of -2 and a mileage demand of -1.

Suppose that t falls when demands and supplies in town have zero elasticities. Areas devoted to the various activities must remain the same, hence distances must remain the same. The distance formula shows that all V's must decrease in proportion to the fall in t. Hence elasticity of demand for transportation is zero, and all rents simply fall in proportion.

Suppose we compare two contemporaneous cities with different populations. Assume, to a first approximation, that total demands and supplies of these cities vis-a-vis their hinterlands are proportional to their populations, and that their technologies are the same, and that unit costs of transportation are the same. Then the area devoted to each activity is in proportion to population; therefore the distance of successive rings is proportional to the square-root of population. This implies, from the distance formula, that the V's are proportional to the square-root of population.

V_1, the rent on the innermost activity (perhaps downtown retail trade), may be approximated by the peak land value, which should therefore be proportional to the square-root of population. Inspection indicates that this relation holds, though with considerable variation for individual cities. (HM Bodfish, J Land & Public Utility Economics, 6:270-77 August 1930; JQ Stewart and W Warntz, J Regional Science, 1:99-123 Summer 1958).

Many of the conclusions remain valid if the assumptions of the model are relaxed. For example, suppose the city is the hub of a system of highways which radiate out in various directions; it may also have a railroad system; it may also lie athwart a river or canal; it may border a sea-or lake-coast; it may have hilly land in one direction out of town and flat land in another. Unit transport costs clearly will vary with direction from town; but suppose that for any given direction transport costs are proportional to distance (this

will happen if all inhomogeneities are on straight lines or sectors radiating from town). Then it may be shown that all previous conclusions, except one, remain valid. The one exception is that borderlines, instead of being concentric circles about the city, will be star-shaped with outward projections along transport arteries. This is the pattern usually observed. (RM Hurd, Principles of City Land Values, 1924).

Again, suppose that transport costs per ton, instead of being proportional to distance, are some general function of distance f (td). (The previous case corresponds to f (x) =x.). Then all conclusions concerning the relations of the C's and V's remain valid, as do the relation of these to rent (but not distance). Elasticity of demand for transportation is still -3 with constant V's (the price of transportation being understood as t in the above function). (On the last two paragraphs cf. MJ Beckmann, Weltwirtschaftliches Archiv, Bd69Heft 2:199-213, 1952).

We may hope to apply the Thünen model to a diversity of situations:the relationship of land-use zones to major population concentrations, such as the industrial belts of North America and Western Europe;the relationship of hinterlands to metropolitan centers; the relation of suburban communities to their central cities; the pattern of land-use zones within cities themselves. It is sometimes possible to estimate C values for different activities and to check predicted ordering (e.g. in agriculture outward from truck farming to sheep ranges).

To illustrate some of the problems which may arise let us take a rather difficult example: the distribution of residential land graded from low-class to high-class. Everyday observation, and also surveys, inform us that grade of residence rises with increasing distance from the center of the city--to be sure, with numerous exceptions, such as Park Avenue, but the general direction of the relationship is clear. If the Thunen model is applicable, this means that high-class residential housing should have a lower location weight per acre, C, than low-class. The activities carried on are sleeping, child-rearing, and family living in general. It is well-known that low-class housing involves on the average more people per acre, both in the sense of having fewer square feet of floor-space per person, and having a greater proportion of multiple story dwellings; also there is less open area per person in the form of lawns, playgrounds, etc. But the conclusion is not altogether clear. Rich people "weigh" more than poor people, because of opportunity costs in traveling; also they consume more per person. These factors work in the opposite direction. Differential ownership of automobiles affects travel costs per person-mile, hence weight per person, hence weight per acre--but in what direction I can't say. We must also consider the differential distribution of trips per unit time, due to varying proportions of people in the labor force and going to school, and varying habits in the frequency of use of entertainment, religious and other facilities. A further complication is that trips are not always, or even usually, on a line with the center of town. It is not clear to me how far this fact invalidates the analysis. Furthermore, there will be a tendency for people with similar tastes--in the same income-class, say--to conglomerate irrespective of their distance from the center of town, since this will allow the advantageous location of special facilities catering to these tastes. (Also, there is an amenity value for rich people, or those of the proper race or religion, to associate with each other. This and the previous consideration belong under the heading of scale economies and external effects, to be discussed below.) Finally, there is an essentially dynamic element which has so far eluded our analysis entirely: cities grow from the center outward, so that on the average the inner houses will be older than the outer; this in itself would tend to produce the observed pattern until such time as extensive rebuilding had obliterated the original age-differential. This may last a long time since people will be loath to build a

new house in a dilapidated neighborhood--(external effects. Does this explain slums?)

Economies of scale and external effects

There still appears to be one crucial unexplained fact in the Thunen scheme: why the city itself exists. It functions here as the distributing center for the economy, and, in the general case, may also contain activities of its own. We may eliminate conceptually this second function by assuming that if such activities exist they also will fall into concentric zones inside the innermost hinterland ring. The city now remains a pure distributing center. Suppose now that the economy splits into two equal pieces, each with its own city and identical concentric rings, and far enough apart so as not to influence each other. Each piece carries half as much tonnage as the original; the area covered is therefore cut in half, and the distance to the corresponding ring reduced by a factor $2\frac{1}{2}$. While total tonnage for both pieces combined remains the same, distance moved, hence transport costs, have been reduced by a factor of $2\frac{1}{2}$. Similarly, if the original economy is split into N equal pieces, transport costs will have been reduced by a factor of $N^{\frac{1}{2}}$. The optimal solution, therefore, would be to split up the economy indefinitely, ending with a homogeneous mixture of activities uniformly distributed over the plain, with transportation costs equal to zero!

There are two reasons why this denoument does not occur. One is the unequal distribution of geographical features. Cursory investigation indicates that much of the gross distribution of population might be explained by this circumstance alone: there is a strong connection between the occurrence of dense populations and good climate, adequate water supply, flat terrain, fertile soil, access to water transport, and, latterly, to coal and iron ore deposits. One might thus be led to predict the concentration of populations in large regions, or along rivers and coastlines--but not their further articulation into cities.

(It is not even clear that homogeneous natural conditions must lead to homogeneous population distribution from considerations introduced thus far. Suppose Adam and Eve were set down in the middle of a homogeneous plain. Would their descendants tend to spread themselves evenly over the plain? Not necessarily: early capital investments--perhaps drainage and leveling of the soil, clearing of forests, construction of houses, roads and wells--would be located near the origin of the pair. But once made, these investments become part of the natural landscape and destroy its pristine uniformity. Whether this initial uneveness would eventually be obliterated or not is a very difficult dynamic question.) On the other hand if the human species were evolving and redistributing itself to adapt to slow geological changes, it is hard to see how the initial situation could arise. This suggests that we look elsewhere to explain the existence of cities.)

The unevenness of natural features (and the unevenness of the distribution of population and capital resulting therefrom) supplies the basis for trade. This implies both regional specialization and expenditures for transportation. Let us concentrate first on the latter; we may conveniently refer to the breakdown of transportation expenses listed above. As for the channel(s) there will usually be a saving in funneling trade through a small number of major transport routes instead of having a separate route for every pair of trading units. (AC Pigou Economics of Stationary States). There will also be an advantage in having a small number of central depots in order to save buyers and sellers the costs of searching each other out.

Other informational aspects of cost reduction through centralization are the easier provision of financial services, the details involved in transfer-of-title, inspection of merchandise, insurance, the setting of specifications from buyers to sellers and price information from sellers to buyers. In addition, there are economies in the bulk handling of goods in facilities for storage, loading and unloading, and transferring goods from one mode of transport to another.

Once a center has come into existence in response to these forces, others arise to accentuate the trend. Situated as it is at the convergence of several transport routes, it is likely that the center will provide a local maximum for the V values of many activities: firstly, because of the relatively low transport costs on inputs and outputs; and secondly, because by locating near a transport terminal an extra loading and unloading cost is avoided. Thus many non-commercial activities will be attracted to the depot, which fact will react again in a beneficent circle on the bulk handling economies of the depot. Finally, a number of activities which are linked to the original settlers will find it advantageous to settle in the center in order to be close to their sources and/or markets. What limits the process? (a) The competition for space near the transport terminal forces activities to choose between paying an ever-higher rent, or locating ever-further away from the center, thus neutralizing the original advantage; (b) increased congestion and crowding and other "external diseconomies" underlines this effect. Nonetheless, the process can be extended, perhaps indefinitely, by the construction of local transport facilities radiating out from the main center to newly formed sub-centers which repeat the process. (This is one aspect of "central-place" theory which will be discussed below.)

The entire development described above required for its impetus the unequal original distribution of natural features. In particular it does not explain how Thunen's isolated state could exist on its homogeneous plain. There seem to be other agglomerating forces at work.

Before we discuss these, it is well to take a harder look at our central concepts of "economies of scale" and its associate "external effects". The two concepts are distinct: one may have external effects without scale economies--e.g. in ordinary monopoly analysis; conversely, one may have scale economies without external effects--e.g. when an exact Pigou-Kahn tax-bounty scheme is in effect. Let us now give a twist to the useful distinction between "pecuniary" and "technical" external effects (J Viner, Zeits, f. Nationalokomie, 1931; cf. JE Meade, Economic J, 62:54-67 March 1952). We ignore the former as irrelevant for our purposes. The latter we define as follows: Consider two activities, each occupying a certain portion of space and a certain interval of time; if there is a non-zero marginal rate of substitution between one or more of the inputs or outputs of one and the other we shall say there is an external effect between them. The effects may be one-way or they may be reciprocal. The externality of the effect is physical--operating where it is not located--rather than proprietal-- affecting someone else's equity (without compensation). The claim now is that all the phenomena subsumed under the rubric "economies of scale" are reducible to these technical external effects.

We first define "economies of scale" to make it symmetric with respect to inputs, outputs and capital tied up. A transformation function is subject to increasing (decreasing) returns to scale at a given feasible combination of inputs, outputs and stocks if with a small proportional increase in all but one of the quantities, the last quantity can (must) increase more (less) than proportionally if an output, or increase less (more) than proportionally if an input or a stock. (Smoothness of the transformation function insures the consistency of this definition.)

We know ecomonies of scale may exist at the level of the firm (more particularly, of the plant) and of the industry; also, they may exist at super industrial levels, e. g. manufacturing in general, or even the entire economy. (These super-industrial, or non-Marshallian economies should perhaps be credited to the discovery of Allyn Young, Economic J, 38:527-42, Dec. 1928, or even to Friedrich List). They may also exist at a hierarchy of regional levels, from the lot, to the neighborhood, to the city, to the metropolitan area, to the nation, to the world. Perhaps a hierarchy of time-intervals would round off the list (successive "long-run" adaptations).

Again, the concept of "scale" as a size measure is ambigous, since it has at least three dimensions: density, area and duration. By an increase in scale we may mean an increase in density, holding area and duration constant, or an increase in area, or an increase in duration, or some combination of these; and each different meaning is likely to have different effects.

The claim regarding external effects may be restated to the effect that the appearance of economies of scale at a certain level of aggregation indicates that there are (good) external effects among its component processes.

Consider first an individual worker, or piece of equipment or structure which is subject to qualitative variation--degree of strength, or skill, or size. It may be, with reference to a given productive process, that a certain threshold level of quality is needed to perform--e.g. a man is capable of lifting a certain weight or not--or there may be a continually increased response in performance--e.g. the capacity of a warehouse goes up as the cube of its linear dimension, while construction costs may rise about as the square. In the second case the greater the overall output the greater the saving, while in the first case it is more a question of getting an exact multiple of the optimal threshold size.

The economical size of such units will depend on the extent of the market for its outputs and of the supplies of its inputs. This will involve all three dimensions--density, area and duration. The presence of abundant rainfall and great city size will allow economies to be realized in the water supply of a city; similarly with electricity, providing coal does not have to be imported from too dispersed a range. Heavy traffic will allow economies in the use of mass transportation equipment and structures, as has already been mentioned.

The gathering together of several industries has the effect of widening the market for each; the resulting expansion provides still further stimulation in a beneficent circle.

A related phenomenon is that of "pooling" or the principle of massed reserves (PS Florence, The Logic of Industrial Organization, Chap. 1). An activity finds itself with a pool of suppliers of its inputs and of customers for its outputs, and therefore saves the need for searching them out, as has already been mentioned. Firms find a pool of available labor, while workers find a pool of available jobs; retail stores find a pool of customers, and vice versa. Auxiliary services, such as banking, insurance, accounting, law, repairs and maintenance, find a sufficient market to exist in quantity.

Massing allows more room for the dovetailing of activities so as to keep factors more fully occupied--e.g. multiple shifts, seasonal interlocking, industries utilizing women and by-products.

A number of other phenomena relate to information. Innovations diffuse rapidly, and perhaps stimulate new innovations more rapidly. The coordination of linked activities becomes easier (though the need for coordination is a diseconomic effect of heavy reliance on the market. Coodination involves scheduling, so that complementary factors can meet without waiting too long for each other; also, so that dead time can be dovetailed and so minimized. Also, it involves standardization, or the optimal adjustment of qualitative variability in complementary factors--e.g. having nuts fit bolts, having a single language, single standards of money, weights and measures.

The variety of products and processes makes finer adaptation possible--e.g. allowing workers to make fuller use of their native capacities, as Marshall has emphasized.

Though this is by no means a complete listing, nor a very satisfactory one, it probably covers the main descriptive forms in which economies of scale manifest themselves. The task now is to estimate their importance in the real world--in particular, among communities of different sizes.

The major forms of diseconomies seem much easier to list. One is congestion: things getting in each others' way. In general congestion will exist on any land whose rent is greater than zero, being carried to the point where the marginal increase in costs due to congestion equals the marginal saving of rent from the compression of space(in the absence of pecuniary external effects).

A second diseconomy might be called "pollution", the term being used in the general sense of the incidentals of one activity interfering with the workings of neighboring activities. Examples are noise, dirt, smoke, smells, eyesores, shutting out of sunlight, "bad" neighbors, traffic hazards, destruction of natural environment; also, from a slightly different angle, danger from contagion and from fire.

A third group are the obverses of the economies of specialization discussed above: lack of adaptability, one-sided development, loss of organic rhythm, insecurity, reduction of human contacts to a money-nexus, are some of the criticisms made by numerous writers. It should be stressed that economies of scale--especially when "scale" means increased density--will always be accompanied by diseconomies; it is the net result which decides.

Central-place theory

Let us return once more to our homogeneous plain and ask what effects the considerations of the last section will have had on population structure. The balancing of economies of scale against increasing transport costs with size of place will produce an optimum size for the city. This size will be higher the lower unit transport costs are. We should then expect a partitioning of the plain into a number of equal areas, each with its nuclear city and surrounded by its own little system of rings. There will be no trade among cities because they have identical structures; there will be trade only between a city and its hinterland.

Each hinterland itself, however, may be affected by agglomerating forces. In the first place, there must be a transportation system between hinterland and city, and this will lead to the formation of a system of channels and depots, as discussed previously. This will distort the circular ring structure: firstly, into the star-shaped pattern mentioned previously; secondly, into a tendency for sub-rings to form around the depots. In the second place, once depots

exist there is no need for all trade to go to and from the central city. Some kinds of circulation may be short-circuited through the sub-centers; only those goods which have to be in-gathered or dispersed over a wide-ranging area will have to pass through the central city. Also, of course, the inputs and outputs of the activities in which the central city specializes will be traded by the city. Thirdly, if economies of scale in one (or a combination of several) of the hinterland activities exist, this will provide still a further force for the concentration of these activities in the depots, or urban sub-centers. Historically, in fact, these sub-centers may first have come into existence through economies of production, and only later acquired their distributional functions.

The picture that emerges is of a central city and several satellite cities at varying distances from it, the whole connected with a web of transportation channels. The sub-centers serve as distributing points for the major center, and also will serve as centers themselves for local trade. All these cities specialize in production to a greater or lesser extent. We may expect, in a general way, that the inner cities will tend to specialize in activities which would have occupied the inner Thunen rings, and the outer cities in the outer activities, and in activities linked to these.

The same considerations as applied to the central city and its hinterland may apply to a sub-center and <u>its</u> hinterland. The result will be a further articulation into sub-sub-centers, and so forth. The overall result, then, is that a <u>hierarchy</u> of cities is formed. Such is the result on a uniform plain if the only scale economies which exist occur at the urban level--that is, depend only on overall city-size. Suppose now that super-urban economies exist (e.g. in the diffusion of knowledge or electricity), "scale" referring to density. This will distort the even spacing of cities over the plain, and tend to bring them into regional clusters, the size of the cluster depending on the level up to which the economies obtain. Suppose, conversely, that there are economies from the industrial-urban level--that is, from the expansion of a single industry (or industry-complex) within a city. Then even on a uniform plain there may be a tendency for different cities to specialize in different activities (or activity complexes). This may also produce regional groupings of cities, the regions in toto being identical with each other, but component cities heterogeneous. Or, it may produce superimposed nets of specialized cities (cf. A Losch The Economics of Location, 1954). In either case an inter-urban transport web will be built up, the depots perhaps being already existing cities.

Non-uniformity in natural features will simply accentuate this trend toward urban specialization, but may also add regional specializations on top of it.

These specializing forces may tend to produce cities of different sizes, since for different activities economies will extend up to different sizes. In particular there may be a threshold size associated with any particular activity, also perhaps a threshold size for the market and supply areas for any activity.

The picture that emerges is the central-lace hierarchy developed (in a more rigid geometric fashion) by Walter Christaller (Die Zentrallen Orte in Suddeutschland, 1933), which combines the ideas of the city-hinterland and the size-threshold specialization of cities. There is a hierarchy of cities(in Christaller's scheme, 2 rural plus 7 urban levels; various other numbers in the voluminous literature which has appeared in the wake of his original study), in which a city of a given level performs central functions for a group of surrounding cities of the next lower rank (3 cities for Christaller; up to 10 for others). Any particular service appears at a given threshold level in the hierarchy and at all higher levels.

The center also acts as a distributing center for its sub-cities. It may be said that the threshold concept for services has been confirmed in a rough way by numerous studies, mostly surveys of rural regions; the exception is the existence of specialized manufacturing activities which are tied to local resources such as minerals or skilled labor, or are simply small-townish, such as cotton textile manufacture. On the other hand, no rigid spacing of cities has been found, nor do the concepts of the rank of a city and its span of control as particular integer numbers seem to be more than convenient fictions. (cf., for example, R Vining, Economic Development and Cultural Change, 3:147-195, Jan., 1955).

With a few extra assumptions, the central-place scheme has implications for the distribution of city sizes. Statistics on the latter from many different times and places indicate that the distribution is usually well-described as a Pareto distribution: the number of cities of population size greater than p is cp^{-a}, C and a being constants specific to the time and place in question; a shows some tendency to decrease with time and to approach the value 1, which is about its mode for the various distributions observed (GR Allen, Bull. Oxford U. Inst. Stat., 16:179-89, May-June, 1954; GK Zipf, Human Behavior and the Principle of Least Effort; GK Zipf, National Unity and Disunity, 1941; HW Singer, Economic J, 41:254-63, June, 1936). The following model is due to Martin Beckmann (Economic Development and Cultural Change, 6:243-48, April, 1958; anticipated verbally by EM Hoover, ibid, 3:196-98, Jan., 1955). Curiously, it is identical in form with that adduced by Harold Lydall to explain the Pareto distribution of employment incomes (Econometrica, 27:110-115, Jan. , 1959). We simplify the notation and argument to some extent:
Let s be the span of control, i.e. the number of cities of next lower rank served by a given city; s is, therefore, the ratio between the total number of cities of two successive ranks; if we let r be the rank of a city (counting upward), the total number of cities of rank r is hs^{-r}, h a constant; the total number of cities of all ranks above r is then the sum of a geometric series, and comes out to be $hs^{-r}/(s-1)$; call this #. We now assume that the populations of cities of successive ranks rises in constant ratio, q, say, so that the population of a city of rank r is kq^{r}, k a constant; call this p. p and # are both functions of r, and we can eliminate r and solve for # in terms of p. This comes out to be # equals cp^{-a}, where c is a constant involving h, k, q, and s, and a equals (log s)/(log q). Thus we have a Pareto distribution. Furthermore, if s is near q (which means the population of a city equals the combined populations of its subordinate cities approximately) a will be near 1. Also, the central-place scheme may offer an explanation for the secular decline in a(that is, increasing inequality in size distribution). An earlier stage of economic development will have less interregional specialization, and so approach more closely the uniform distribution of equal city-sizes discussed above.

The model is not free from objections, however. In the first place, the values r, and hence s and q seem to be convenient fictions, as mentioned above; also, if ranks existed, it is not easy to see why the s and q values (or rather their ratio s/q) should be constant from level to level. Finally, it does not fit all the facts. The central-place scheme seems to fit the facts best in rural areas (farm trade centers and the like). But the Pareto law breaks down for rural towns: the one piece of evidence I have examined so far (JE Brush, Geographical Review, 43:380-402, 1953), which charts the towns from population size 7000 down to population size 10 in South-west Wisconsin, shows, for towns below 1000 population a beautifully regular but non-Paretian distribution. (The distribution is given by the number of towns in the population stratum from x to y being proportional to log (y/x). So far I can find no explanation for this result, save for a curiosum noted below.

Stochastic models

I have not found any models explicitly tailored to the stochastic explanation of city size. However, the same model-types sometimes apply in widely different contexts; accordingly, I searched around for likely-looking ones, especially such as lead to Pareto distributions. These occur in the distribution of incomes and of word frequencies. Many of the processes which occur in these do not make sense when applied to cities--e.g. inheritance. On the other hand, the model of Simon (Biometrika, 52:425-40, 1955) in some ways applies more plausibly to cities than to the word frequencies for which it was mainly intended. We will present it directly in city-size categories. Imagine a certain distribution of cities to which people are being added one by one; the chance of a person being added to a city of population i is proportional to the combined population of such cities; also there is a fixed chance of the new person founding a new "city". The number of i-cities will be increased if the person lands in an $i-1$ city and decreased if the person lands in an i-city. If $f(i,k)$ is the expected number of i-cities when the total population of all cities is k, and a is the chance of founding a new city, we obtain the equations

$$\text{for } i>1: \quad f(i,k+1) - f(i,k) = \frac{(1-a)}{k}\left[(i-1)f(i-1,k) - i f(i,k)\right], \text{ and}$$

$$\text{for } i=1: \quad f(1,k+1) - f(1,k) = a - \frac{(1-a)}{k} f(1,k).$$

A distribution which satisfies these equations asymptotically when k is large, and is, moreover, stable, is $f(i,k) \approx \frac{k(\rho-1)\rho!\,(i-1)!}{(i+\rho)!}$

where $\rho = 1/(1-a)$. This is known as the Yule distribution. For large values of i it may be shown to be asymptotically Paretian with exponent ρ

In evaluating Simon's model we first note that even if it were satisfactory in all other respects it would still be incomplete, in the sense of leaving out variables which we know to be relevant in the determination of city-size. For example, nothing is said about the region in which a city is located; yet it can be easily verified that the cities of a region as a whole may have abnormally high or low growth rates (in the United States, the Pacific region and New England, respectively). All central-place concepts are absent. The previous rate of growth of a city is not mentioned; yet cities show a high correlation in their growth rates for successive decades (CH Madden, Economic Development and Cultural Change, 6:143-70, Jan., 1958). Cities also tend to go through a cycle of early rapid growth and later retardation as they age (ibid).

Still, we should not ask more of a model than it is designed to give. We may legitimately criticize a model for (a) having unrealistic concepts and postulated relations, and (b) making wrong predictions. Simon's model--at least this simple version of it--is a pure birth process: no provision is made for deaths and migrations. This fact makes it suspect, since migration is easily the major force making for population redistribution; in fact, natural increase tends to vary inversely with overall increase (RB Vance, Research Memorandum on Population Redistribution within the United States, 1938). Deaths could probably be incorporated with little difficulty--the chance of a death being proportional to population size, etc.--and would lead to relatively little modification of the result.

But migration could not realistically be introduced on a similar random basis: there must be a definite bias in favor of migration from smaller to larger places. One could estimate these chances from the known (rather scanty) data, and find out what would happen if this regime operated indefinitely, but this result, while interesting, would leave most of the more intriguing questions

unanswered. A more fruitful approach might be to introduce explicitly at least the distance among the cities, towns and rural places. This would lead to the construction of a "quasi-gravity" model (see below)from which might be coaxed the associations mentioned above with region, previous growth, and age.

It is also clear that any model involving the founding of new cities must include migration to make sense.

A curiosity, which may bear further investigation, relates Simon's model to the anomalous distribution of rural communities mentioned above. If the Pareto exponent is near one, the Yule distribution comes out to be approximately proportional to 1/i (i + 1). If we interpret i to mean, not numbers of people, but numbers of clumps of people, say population in thousands, it becomes meaningful to work with fractional i's. For i well above one ("urban" population) we get the Pareto upper tail of the Yule distribution; for i close to zero ("rural" population) the density function becomes proportional to 1/i, and the number of cities between two sizes proportional to the logarithm of their ratio, just as in the Brush distribution!

The secular decline in the Pareto parameter might be incorporated into Simon's model by its allowing a to decline with time--that is, allowing a decline in the rate at which new cities(farmsteads?) are founded relative to the rate of population growth. This would seem to be a concomitant of economic development.

A word may be said about gravity models. The problem is to predict the intensity of interaction between two places in various respects, such as ton-miles of freight in a certain time period, or number of auto or plane trips, number of telephone calls, amount of migration, and so forth. A large number of studies of this kind have been made, the most popular explanatory equation being of the form: intensity equals $P_o P_d / D^n$, where P_o and P_d are the populations of origin and destination, D is the distance between them, and n is a constant exponent to be determined. Sometimes more sophisticated measures of these size and separation parameters may be used. The P's may represent number of telephones, or total income, or even total interaction with all other places (which loses some degree of freedom); D may be better represented by transportation cost. n may be determined by regressing $I/P_o P_d$, or rather its logarithm, on the logarithm of D (e.g. JD Carroll and HW Bevis, Papers and Proceedings of the Regional Science Association, 3:183-97, 1957). The exponents thus obtained range empirically from about 1 to 3, with a mode about 2. The interesting fact about these models is that with only two adjustable constants (exponent and scale) they sometimes achieve notably close fits: correlations above .9 are not unusual. The rationalization of these models by some kind of probability approach that makes sense in terms of location theory is desirable, but I have so far not found anything satisfactory.