THE ECONOMICS OF URBAN SIZE

William Alonso
Professor of Regional Planning
University of California, Berkeley

November 1970

Working Paper 138

I want to thank the Centre for Environmental Studies, London, for their help.

This work is supported by a grant from the Economic Development Administration, U. S. Department of Commerce.

To appear in Papers of the Regional Science Association, Vol. XXVI, 1970.

At least since Aristotle, men have wondered about the best size for cities. In the last decades developed and developing nations, capitalist and socialist, have increasingly adopted more or less explicit policies on urbanization with special reference to city sizes. Most typically these policies assume that the big cities of the nation are too big, and therefore try to disperse growth. Complementarily, in recent years such dispersal policies, and policies addressed to distressed or backward regions, have recognized that these alternative centers must be of a certain minimum size, however ill-defined, in order to be viable. In its simplest sense the question of urban size consists of symmetric parts: how big is too big? and, how big is big enough?

Theory and fact on these issues are scarce and poor, and they swim in an ocean of opinion, much of it highly emotional. For instance, recent American academic opinion (Berry [6], Thompson [25]) places the minimum size in the area of 250,000 to 500,000 inhabitants, which is about what the Soviet Union regards as the maximum tolerable size. The designation of "growth centers" has been awarded to settlements ranging in population from less than a handful of thousands to more than one million.

Here we will present an aggregative economic approach to the theory of city size, and some empirical findings which suggest that even the largest cities have not yet reached excessive sizes from the point of view of growth and productivity. The discussion of least sizes will be more sketchy. Also, some observations and some data will be offered on the question of the size of a city in the context of its position in a system of cities.

The Difficulty of Defining Urban Size

Urban magnitude is no simple one-dimensional phenomenon. For instance, an Asian metropolis of 5,000,000 inhabitants with a gross regional product per capita of $100.00 has an economic magnitude equivalent to that of a modest American metropolis of 100,000, and is far smaller for income-elastic goods and services. Nor is the definition of population size unequivocal. Modern urban centers are surrounded by very large, diffuse zonal boundaries, within which there are marked variations in the proportion of firms and people associated with that center, and in the intensity of the association. Thus, population does not constitute a conventional countable set, where people are unequivocally members or not. The situation is closer to that of "fuzzy sets," in which an element's membership in a set is a matter of degree (see Zadeh [28]), and so a number as a measure of population is a gross oversimplification. The situation is further complicated in the frequent case of conurbations or megapolitan areas, where the zonal boundaries of diverse centers overlap in complex patterns, and a person may be a member of two functional cities. To avoid the problem of the definition of boundaries (if not that of determining the degree of membership) some scholars have used density rather than size, but density is only a measure of local intensity of a ratio which misses the crucial aspect of the extension (or scale) of the system of interrelated elements. Certain villages of medieval origin have higher densities than giant metropoles or even their central cities.

In general, we will ignore these problems, considering population as the basic magnitude and as a conventionally definable number, although some of the evidence to be cited is based on density measures; and towards the end of the paper I will consider some effects of adjacency of urban

areas, although not the overlapping of their boundaries. Economic magnitude will be treated as an endogenous variable, basically a function of population.

The Minimum Costs Approach

Most approaches to city size have stressed the presumed diseconomies of urban scale, and have sought to establish that population at which costs per capita are least, regarding this as optimal. Although I shall argue that both the logic and the factual basis of this approach are faulty, it will be useful to review them briefly since they are so widely accepted by scholars and policy makers.

Traditional studies of urban economies, going back to the turn of the century, have focused on how public costs vary as a function of population size using cross-section data. In general they have found them U-shaped, with the bottom of the curve occurring variously between 10,000 and 250,000 population. This literature is well known and has been supported by a few studies based on engineering calculations (for useful recent reviews, see Cameron [8], Hirsch [12], Kain [14]). But the matter is not so simple, and these findings cannot be accepted at face value. Three principal difficulties may be mentioned: (a) These measures of cost measure only inputs, and implicitly assume that outputs are constant. If the demand for public goods and services is at all income-elastic, cities with higher incomes would be spending more to get more, so that rising expenditures are not strong evidence of rising expensiveness. Indeed, the few studies based on multivariate techniques find no significant correlation of size and public cost after other variables are taken into account (see, for instance, Schmandt and Stephens [22]). (b) The division between private and public

costs is very much a matter of institutional convention. Most automobiles are private, but buses may be private or public. The production and distribution of electricity may be public or private. The sewer and waterpipes that vein a suburban residential district are public, but their exact equivalents, running vertically within a large apartment house, are private. In brief, the category of public costs is neither well-defined nor stable.

(c) Many of the components of costs may not be real economic costs. For instance, suppose a teacher receives a higher salary in a large city because teachers there are unionized; the difference in his salary represents a transfer payment within the city rather than a true resource cost. It is unclear, in fact, how much of education is a production cost (training people) and how much of it is a form of consumption (educating people). Similarly, the treatment of land costs is ambiguous in the cost-benefit literature.

By contrast to the extensive literature on the costs of infrastructure and municipal operation, there is only a very slender literature on the variation in producers' costs with city size. No general study of the variation in producers' costs with city size appears to exist. A recent study in an Indian context (Morse [17]) finds no substantial variation (and a possible small decline) in a range from very small cities to rather large ones. In the case of consumers' costs, it appears (Alonso and Fajans [4]) that these vary only slightly with urban size. The association is weak even for the housing and transportation components, which from theory might be expected to be strongly associated with urban size. The association disappears if other factors, such as local climate and income are taken into account. Subjective estimates (Gallup [11]) show, however, a sharp rise in the level of income that people think is needed for adequate levels of living

in larger cities. It is popular opinion, of course, that big cities are more expensive. It appears that one can live as cheaply in big cities as in small ones, but that the more varied opportunities of large cities raise expectations.

The most sophisticated explanation of the excessive growth of cities runs as follows: where costs are rising, a new industry (or inhabitant) makes its location decision on the prevailing (average) costs, including such factors as congestion and local taxes. However, since costs are rising, marginal costs are greater than average costs. Marginal costs are borne by the urban body as a whole, and the differences between average and marginal costs are the negative externalities. For instance, a plant considering locating in a large city will take into account existing (average) levels of congestion, but does not consider the increased congestion and travel costs that would be borne by the whole population as a result of its coming. By this argument, then, this divergence between private and social costs permits the city to grow beyond its best size.

It would be clearly impossible in real life to apply to each firm and citizen in order of arrival, such a differential tax, corresponding to the difference between marginal and average costs. Nonetheless, this view is reflected in the tax policies of many countries, such as Great Britain and France, which impose surtaxes on capital and/or labor in locations which are thought congested. Although such taxes operate on the average value for wages, they approximate a marginal approach for capital when they are levied on new investment. In many cases, of course, governments do not trust such uncertain subtleties of pricing, and public policy manifests itself through direct command on industry and population, denying them some locations or even ordering them to others.

An Aggregate Theory of City Size

The argument of minimum costs is insufficient in its own terms. Such an objective is sensible only if output per capita is constant. But, in fact, it appears that output per capita is an increasing function of urban size. In that case, a more sensible objective of public policy would deal with the relation of outputs and inputs, rather than only with inputs.

Before going on to spell out this simple point, we must return to definitional difficulties. It is in many cases most difficult to determine whether something is an input or an output. We have already raised this question with respect to education. But examples abound: for instance, are the expensive tie and suit that a businessman wears to his office production costs, or are they a form of consumption? Such problems, which had troubled early theorists of national income accounting, have been largely ignored by consensus in recent years. Yet they constitute a crucial area of ambiguity in the type of theory with which we are dealing, because they raise fundamental questions as to which human activities are instrumental and which are ends in themselves. For our purposes we will consider that urban output is the value of the total product of the urban area; urban costs are harder to define, and would include quantity and price effects in the costs of infrastructure and municipal operation, in the costs of exogenous inputs other than human ones into the city's economic activity, and in private consumption. Thus, we regard the city as an aggregate productive unit.

Figure 1 shows a possible set of cost and product curves, and is akin to the traditional diagram of costs and revenues for the firm. The key difference is that the horizontal axis is in terms of population rather than of quantity produced. While the usual theory of the firm treats labor

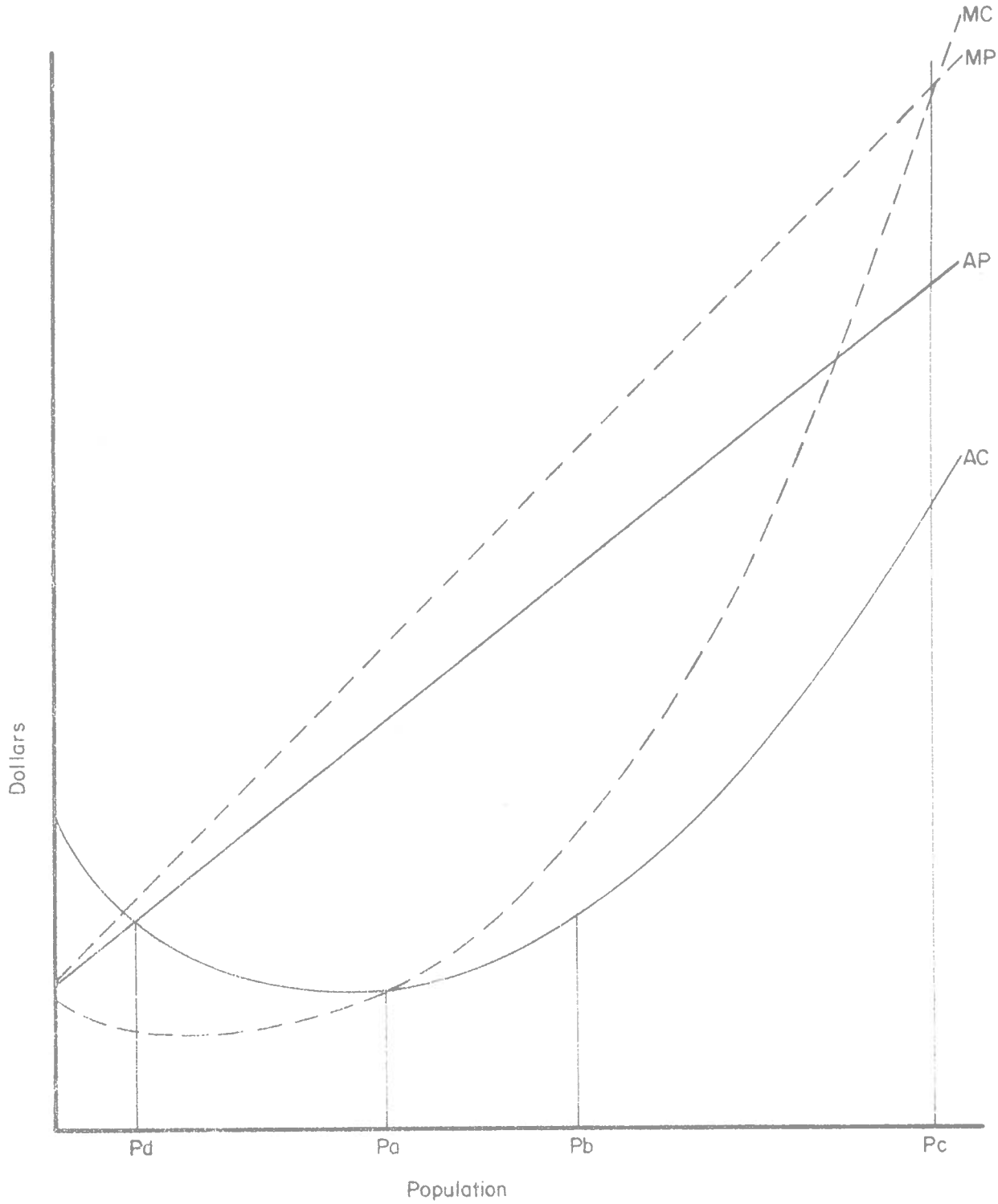

Figure 1: URBAN COST and PRODUCT CURVES with CITY SIZE

as an input whose price is exogenously determined, here labor is excluded from consideration in the construction of the cost curve, and the return to labor (in the broad sense of the total urban population) is the difference between the value of total output and total costs. Here I exclude from consideration colonial situations, in which part of this output is alienated by others elsewhere. The difference between costs and the value of output is thus available to the local population for investment or direct consumption, either directly from the city's production or through trade with others.

The average costs curve (AC) in Figure 1 is shown as rising after a certain population level, both because this is generally believed to be the case (although the factual basis was questioned above) and because it weakens and therefore tests more sharply the argument being presented. The average product per capita curve (AP) is shown rising monotonically, partly to simplify the argument at this stage and partly because this is what most of the empirical evidence suggests. Later on we will consider the possibility of its turning down after some point.

If the general formulation is accepted, the rest is simple. The point of minimum per capita costs, P_a, is uninteresting. The point of maximum local contribution to national income occurs at P_c, where marginal costs (MC) is equal to marginal product (MP). There is of course, no need for this point to exist, in the sense that MP may remain higher than MC for the relevant range of population, but if it does exist, beyond this point further population costs more than it is worth. A national government interested exclusively in maximizing total product under conditions of labor surplus, would use such a population as its target. However, if there is not an unlimited supply of labor, the population size that would

maximize national product would be smaller and would occur where the difference between MP and MC is equal to the opportunity costs similarly defined at alternative locations. From the point of view of the inhabitants of the city, however, a more sensible objective would be the maximization of the difference between average product (AP) and average costs (AC). This difference may be regarded as a per capita "disposable income," and it will be maximized, of course, where the rate of increase of AP and AC are equal (dAP/dP - dAC/dP). Since average product is increasing with population, this must occur at a size (P_b) greater than that of minimum costs. In brief, the optimal population will differ according to whether a national or a local viewpoint is assumed, but in neither case will it coincide with the point of minimum costs.

From the point of view of policy, in trying to bring private costs and benefits into line with social costs and benefits, the same logic that leads minimum-cost theorists to argue for taxes on new arrivals based on the difference between social (MC) and private costs (AC), would argue here as well for subsidies based on the positive externalities produced by the new arrival. Thus, the net tax for subsidy would be based on a calculation of (MP-AP) - (MC-AC), which may be positive or negative. Needless to say, these numbers are very elusive ones, and no operational tax could be based on such arithmetic.

There is no attempt here to develop a theory to explain the increase in production with population size. Pieces of such a theory exist in the literature (Thompson [24], Alonso [2], [3]) and they stress economies of scale, advantages of specialization, agglomeration economies, probabilistic needs for lower relative reserves of inventories and labor, ease of communication, the richness of opportunities and the adaptability of large and

complex systems, and so forth. Within the dishevelled theory of externalities, they suggest that positive external economies exceed the negative ones, principally by modifying the production functions of the component activities so that inputs are used more effectively. Without trying to formulate a general theory of causes, we pass to an examination of the empirical evidence.

Empirical Evidence

In every country for which I have found evidence, local product per capita (or some index for it, such as income or wages) rises with urban size, and where comparable figures on cost are available, these rise far more slowly if at all. Although all of the data one might wish for is not available for any single country, the overall pattern is clear.

In some ways, the best figures available are those for the German Federal Republic and for Japan. West Germany offers estimates of Gross Community Product per capita, as shown in Table 1, which rises by 40% from the smaller cities (20,000 to 50,000) to those above 500,000. Although public expenditures per capita rise by 44%, the excess of product over expenditures rises by over 30% from the smallest to the largest size class.

The data for Japan, developed by Mera [16] and shown in Table 2, are organized by density by prefectures, with the densest areas corresponding to the largest cities. Income per capita (column 1) rises smoothly by nearly 80%. While the curve of government expenditures is U-shaped (column 2), it is very shallow, partly because of the centralized nature of the Japanese system which, according to Mera, provides for quite a uniform level of public services throughout the nation. This would eliminate variations in these costs that arise from higher levels of demand generated by higher

TABLE 1

Gross Community Product and Public Expenditures Per Capita, German Federal Republic, 1964

Population Size Class	Gross Community Product per Capita (DM) (1)	Public Expenditures per Capita (DM) (2)	Public Expenditures as % of G.C.P. (2)/(1)	Excess of G.C.P. over Public Expenditures (1)-(2)
20,000 - 50,000	5,400	513	9.5	4,900
50,000 - 100,000	6,000	627	10.5	5,400
100,000 - 200,000	5,900	603	10.2	5,300
200,000 - 500,000	6,400			
500,000 - 1,000,000	7,500	738	10.4 (a)	6,400 (a)
1,000,000 - and over	7,500			

(a) The arithmetic mean of the figures in (1) was used for this composite class.

Source: Based on figures from Statistisches Jahrbuch Deutschen Gemeinden (52), Braunschweig: Waisenhaus Buchdrukerei, 1964.

TABLE 2

Japanese Population Density, Per Capita Income, Governmental Expenditures, Social Overhead Capital Stocks (SOCS) and All Government Investment Expenditure by Prefectures (1000 Yen)

Population Density (persons per sq. km.)	Number of Prefectures	Mean Per Capita Income (1965) (1)	Mean Per Capita Government Expenditure (1965) (2)	Mean Per Capita SOCS (1963) (3)	Mean Per Capita All Government Investment Expenditure (1966) (4)
Less than 200	17	188.0	50.8	268.5	32.8
200 - 300	15	197.6	46.7	244.8	27.3
300 - 600	8	209.1	43.2	206.4	27.6
600 - 1,000	3	228.0	41.9	205.3	27.5
1,000 - 3,000	1	280.0	53.7	178.4	40.5
3,000 and over	2	340.0	56.1	188.0	39.4

Note: 1960 Population in all cases.

Source: Adapted from K. Mera, "On the Concentration of Urbanization and Economic Efficiency," Economics Department Working Paper No. 74, International Bank for Reconstruction and Development, Washington, D.C., 1970.

income, and leave only technological and price effects. Indeed, the Japanese relative variation in per capita government expenditure from highest to lowest is only about half that of the American or the German equivalent.

Perhaps the most surprising element in the data is the sharp decline with increasing density in Social Overhead Capital Stocks (SOCS) per capita (column 3). This runs counter to common belief that the bigger the city the more infrastructure per capita is needed, and may be the result of such effects as the greater linear quantities of roads per capita necessary in low density areas. The lowest SOCS per capita are for the penultimate density class, but this class consists of a single prefecture adjacent to Tokyo, and presumably shares some of the advantages of Tokyo's size. Yearly government investment expenditures (column 4), which may be interpreted as marginal SOCS, are also U-shaped, declining at first and rising for the densest two classes. However, such investment may be expected to be more a function of the rate of population growth than of the actual density, and the most urbanized areas are receiving a greater share of immigrants.

Turning now to American data, Fuchs found a steady rise of about 40% in hourly earnings with urban size after discounting the effects of labor composition by color, age, sex, and education (see Table 3). If, by traditional economic theory we equate wages to the marginal productivity of labor, the conclusion is that, for a given type and quality of labor, the rise in wages indicates that productivity rises with urban size.

Comparable data, standardized by industrial composition rather than by the characteristics of workers, is shown in Table 4. Although allowance must be made for the difference in the range and disaggregation of the population classes, the results are fairly similar to Fuchs'. The Payroll per

TABLE 3

Ratio of Actual to "Expected" Hourly Earnings, by City Size, United States, 1959

	White Males	White Females	Non-white Males	Non-white Females
Rural	0.83	0.83	0.78	0.76
Urban Places				
Under 10,000	0.84	0.84	0.75	0.63
10,000 - 99,999	0.91	0.88	0.76	0.78
Standard Metropolitan Statistical Areas				
Under 250,000	0.97	0.94	0.84	0.76
250,000 - 499,999	0.97	0.96	1.04	0.90
500,000 - 999,999	1.02	1.03	1.07	1.00
1,000,000 - and over	1.12	1.13	1.10	1.19

Note: "Expected" hourly earnings were calculated for each sex and color by multiplying the national average hourly earnings of each color and sex by age and education by the annual hours worked by members of that cell in the region, summary in each case across all such cells in the region and dividing by the total man-hours of the region.

Source: Adapted from V. Fuchs, Differentials in Hourly Earnings by Region and City Size, National Bureau of Economic Research Occasional Paper No. 101, New York, 1967.

TABLE 4

Indices of Value Added, Payroll, Value Added, and New Capital, by Employee, Standardized by 2-digit S.I.C., 67 Largest Standard Metropolitan Statistical Areas, United States, 1963

Population	Payroll per Employee Index (1)	Value Added per Employee Index (2)	Value Added minus Payroll per Employee Index (3)	New Capital per Employee Index (4)
250,000 - 500,000	0.994	0.970	0.943	0.967
500,000 - 1,000,000	1.029	1.047	1.099	1.028
1,000,000 - 5,000,000	1.061	1.046	1.029	0.925
5,000,000 - and over	1.058	1.053	1.129	0.784

Source: Based on data in R. Douglas, "Selected Indices of Industrial Characteristics for U.S. Standard Metropolitan Statistical Areas, 1963," Discussion Paper No. 20, Regional Science Research Institute, Philadelphia, 1967.

Employee Index (column 1), rises with urban size, although it flattens over the last two classes. Value Added per Employee (column 2), rises a bit more strongly. Taking the wage component out of value added, the remainder of value added (column 3) shows a very strong uninterrupted rise which is the more remarkable in view of the decline of new capital per employee (column 4) (information on capital stocks is not available). Thus, the rise in value added per worker can be attributed neither to rising wages nor to more massive use of capital, and may be attributed to the positive externalities of urban size. We shall return to this below.

Many studies over the years have established that in the United States, for data grouped by city size, there is a strong and steady rise in income, of about 30% depending on the size of classes and income definition used. Sociologists who have looked into various correlates of urban size conclude: "Of all the differences among communities of different size revealed in this study, perhaps the most striking is the pronounced direct relationship between size of place and income" (Schnore [23]). However, since group data tends to subsume intra-class variance, I have experimented with regressions of income on population (Alonso and Fajans [4]). A simple logarithmic regression (not reported in that paper) metropolitan population accounts for 42% of the variance in mean per capita incomes. This is indeed remarkable if one considers the great diversity among cities of climate, resource endowments and history. The relation is only slightly diminished when metropolitan incomes are deflated by local cost of living in the smaller set of metropolitan areas for which the information is available. While local government expenditures per capita rise from $120 at population 15,000 to 50,000 to $200 in counties of more than 1,000,000 population (Schmandt and Stephens [22]), this rise, which represents $250 per family, is made insignificant by a rise of $1500 in income over the same range.

Higher incomes are found in bigger places in developing countries and in socialist countries as well, although there are significant understandable exceptions. Thus, the per capita income of Rome is lower than that of the great industrial cities of the Italian north, and in some developing countries there are some steel or oil cities where incomes are exceptionally high. Similarly, Birmingham, Alabama, a steel town, has a higher payroll per employee than does New York. Particular circumstances are always at work, and here we are speaking of general patterns which are useful for policy analysis, rather than for particular project evaluation.

Similar questions are being raised in the Soviet Union. V. Perevedentsev [19] observes that the extensive Soviet literature on city size, while speaking of "excessive growth," "excessive concentration," and the like, offers no criteria for determining what is excessive. He notes the continued growth of larger cities, in spite of strong direct controls on personal and industrial location to stop this growth, while the productivity of labor is 38% higher and the return on assets is more than twice as high in cities of more than 1,000,000 than in cities of between 100,000 and 200,000. In contrast to these facts he observes that "according to the prevailing views of our city planning, cities with a population of 50,000 to 200,000 are considered optimal, and those with up to 400,000 are permissible." In a reply, B. Khorev [15] speaks of the particularity of cities, especially in their industrial composition, of the higher costs of infrastructure in large cities, and of the value of the additional time spent in daily commuting. He shifts the question of optimal size to one of the particular circumstances of a city and to the hierarchy of cities, but does not offer operational guidelines or systematic documentation.

Since we conclude that it appears that the biggest cities are not too big from the viewpoint of economic efficiency, it may be asked whether the higher average incomes of bigger cities do not mask sharper inequalities among their citizens, so that efficiency is gained at the cost of equity. This does not seem to be the case, at least for the United States. On the contrary, some recent studies (Ornati [18], Burns [7]), indicate that there is less poverty and a more equal distribution of incomes in big cities than in smaller ones. In the pages below, I will touch on the relation of urban sizes and interregional disparities.

A Possible Turndown in the Product Curve

A question of particular interest, even if the general rise of income with size is granted, is whether there is a turndown in the relation after some size. I have tried very hard to test for this with income data for the United States, and in some cases equations using the equivalent of a polynomial of population (i.e., population and its natural logarithm) do yield such a turndown, somewhere between 3,500,000 and 9,000,000 population. However, the standard error of estimate as to where the inflection point occurs is extremely high, and not even New York can be excluded with any certainty from the rising trend, even taking into account differences in cost of living (Alonso and Fajans [4]).

Nonetheless, the possibility of such a turndown is of extreme importance, for it would suggest that those cities were in some sense too big for that society at that time. Therefore, the question must be asked: assuming that a turndown does exist, however weak its indication, does this mean that cities larger than the inflection point are too large? Three reasons suggest themselves why this might not be the case:

(1) The development of very dispersed patterns of suburbanization far beyond the territorial definition of the standard metropolitan statistical area, sometimes called exurban is commonly observed in the United States and has been recently strikingly documented by Berry [6]. These exurbanites typically enjoy greater incomes than the mean of the SMSA population, and are not included in the statistics for the SMSA. If the relative importance of this phenomenon is larger for larger areas, as anecdotical evidence suggests, the mean income of the larger areas would be understated, and the turndown would be more apparent than real. However, it is no easy matter to test for this, for this dispersed metropolitan population is only a fraction of the population resident in the surroundings of the SMSA, and because the largest metropolitan areas are set in a context of other nearby metropolitan areas whose fuzzy zonal boundaries overlap.

(2) On the other hand, the decline might be real and yet not signify that these largest cities are too large. They have traditionally performed a port of entry function for foreign and agricultural populations which have low productivity and earning power during the years and even generations of their acculturation. As urbanizing centers they perform a systemic function for the nation as a whole, and other cities benefit in time as the acculturated citizens disperse to other centers. Inclusion of a nonwhite variable, as a rough test of this hypothesis, weakens or eliminates the turning down of the curve.

(3) Lower incomes and lower wages in the largest cities could be consistent with their efficient size in a nation with a high rate of innovation. It has been argued by several authors in recent years (Hirschman [13], Vernon [26], Alonso [2], [3], Thompson [24], [25]), that the largest

cities serve as seedbeds for new economic activity at a time when its technology is in rapid flux, the nature and extent of its demand unknown, and its institutional structure unformed. During these formative years such new activities depend on an environment rich in external support, where there is available direct access to customers, financing, shipping, jobbing, specialized labor, rumors, ideas, and a thousand other things. If the enterprise grows and succeeds, its management, marketing, and production become routinized into well-defined roles, its technology and demand patterns stabilize so that it becomes possible to standardize and routinize, and hence to substitute capital for labor. With size and predictability, many of the externalities can be internalized. At this point, industries may migrate to provincial centers (often through branch plants) seeking advantages of transportation costs, of cheaper and plentiful labor, of particular linkages with suppliers, or of economies in other inputs or distributional advantages. This pattern might be expected to lead to lower earnings in the centers of innovation because the fluidity of infant industries logically leads to low capitalization, and from this to a lower marginal productivity of labor, and therefore lower wages. This evolution has often been described and documented anecdotically. The most general historical documentation is that by Pred [20]. The cross-sectional data of Table 4 supports this view. The rise in Payroll per Employee suggests that the lower wages arising from lower capital inputs are to some degree compensated by the greater productivity of the larger cities. But the Value Added minus Payroll per Employee column, which discounts wages and rises sharply with urban size, stands in sharp contrast with the overall decline in the New Capital per Employee Index column. In a country such as the United States, where capital markets are relatively well integrated

it is not credible that there should be such sharp differences in returns to capital at different locations as these two columns would indicate. Rather, after allowance of returns to capital, it would appear that there are strong returns to a factor that is often neglected in this type of analysis because it is hard to measure: This is return to entrepreneurship, or entrepreneurial rent and wages. Within this I would include the return to institutional forms when these constitute a local resource. Thus the data in Table 4 are perfectly consistent with this theory although, of course, they do not prove it. It would be possible to have slightly lower income or wages for the largest cities, consistently with overall national efficiency on the basis of the dynamic and innovative systemic function of these centers.

To summarize: there are some weak indications of a turn-down in the product curve at the largest urban sizes. However, even should the turn-down be real, this would not be inconsistent with the efficiency of those larger sizes in a hierarchical system of cities.

The Minimum Size of Cities

Many countries are today engaged in policies for the promotion of economic growth in distressed regions and in frontier regions. Although there seems to be a general agreement in the use of growth centers to concentrate development efforts, the theoretical and factual bases are even weaker for small cities than for large. Economic base, input-output, and other multiplier theories are insensitive to scale effects and to positive or negative externalities, and most other theory seems to hover somewhere between intuition and poetry, with reference to thresholds, will to develop, industrial vocation, propulsive industries to serve as the engine for development, and the like.

This uncertainty as to size is illustrated by a range of three orders of magnitude in the population of growth centers from one country to another. Not surprisingly, empirical studies are extremely scarce since there is lacking a theoretical foundation on which to base factual questions. A few recent studies in the United States have hazarded guesses about a minimum population size for self-sustaining growth. Berry [6], on the basis of a break in the rank-size relation, suggests the quarter million mark. Thompson [24], on the basis of recent growth statistics, places the level somewhat higher. Alonso and Medrich [5], also on the basis of historic growth patterns, identify spontaneous growth centers throughout the size hierarchy. It seems unlikely that there would be a well-defined threshold. Small economies may be expected to be much more particularistic than large ones, where something similar to a law of large numbers might be expected to operate. In a small economy, where the elements are few and the connectivity of linkages therefore relatively low, particular events and circumstances may be expected to play a larger role. The fortunes of a particular corporation, or the sudden economic relevance of a local resource or locational advantage can produce strong rates of growth or decline in a small economy, while in a large one there is always a very strong probability that compensating forces will dampen fluctuations. Although there are strong overall patterns in the rates of growth by urban size, the smaller centers exhibit a far greater variation in their rates.

The greater particularity of smaller centers and the resulting greater variance in their economic and demographic rates, make an aggregative approach such as that presented below of relative little usefulness for the consideration of any one particular small center. Such an aggregative

approach is meant to shed light on general policy rather than on a particular program of project evaluation.

If, as suggested, the firm perceives its costs and revenues as average costs and average product, examination of the left side of Figure 1 shows clearly why a firm might find disadvantageous location in a small city while, at the same time, such a location might be socially valuable. To simplify the argument, consider for the moment that wages are included in the costs. At populations smaller than P_d, a firm would be, in fact, losing money; a location would not become attractive until the difference between AP and AC became comparable to the firm's opportunity costs at other locations. The threshold of spontaneous self-sustaining growth would occur at that point. Nonetheless, if productivity increases with size, marginal product will be higher from the social point of view than average product even at small sizes; and if costs are declining, social marginal costs will be lower than the costs as perceived by the firm. Therefore, it might make excellent sense to subsidize such a firm, either directly or through infrastructure, to induce it to locate in the growth center. This is particularly so if society has a longer range view and is considering a substantial increment in size from the location of several firms, while the rise from the location of one firm would be too small. The great difference is that in large cities the external benefits foregone by the firm (the difference between marginal and average product) are compensated by the implicit subsidy of its paying average costs, which are lower than marginal costs (we are assuming that costs are rising). In the smaller city, the differences between average and marginal rates of both cost and product operate to the disadvantage of the firm.

Such a subsidy to small cities would be justified from the point of view of national efficiency only if the opportunity costs, again defined as the difference between the marginal rates, were lower in the large cities than in the proposed growth center. Our empirical evidence suggests that this is seldom the case. Only in an economy with labor surplus, if capital costs are included within the cost curves, would there be no opportunity costs to the location of population, and would a general policy of development of smaller centers be consistent with a national goal of efficiency. But such surplus populations are common only in developing countries, and in these the scarcity of innovational, managerial, and institutional resources make unrealistic such an urbanization policy.

The development of smaller centers in distressed areas is more likely to find its justification in equity or distributional objectives. A small center might be aided to the point of self-sustaining growth for the purpose of improving local levels of welfare, but this will typically be done at some cost to national product. For instance, Mera on the basis of similar reasoning estimates that the equalization of per capita incomes by prefectures in Japan would reduce national income by 15 to 30%. While the objective of equalization of interregional levels of welfare is a widely accepted fact, much national policy (such as that of the United States) asks simultaneously for conflicting goals: the minimizing of disparities together with the maximization of output.

A similar logical contradiction occurs frequently in attacks on large cities. They are charged at one and the same time with (1) being grossly inefficient, in that increases in population cost more than they produce, and (2) increasing interregional disparities in levels of welfare through

development is observable in most developed countries, where many of the fastest growing metropolitan areas are the smaller ones within the megapolitan complexes (Alonso and Medrich [5]).

It is a pity that the term megalopolis has acquired negative connotations in journalistic usage. There is a popular image of a teeming anthill, of urban areas running into each other and choking up open space. This is misleading. For instance, the first-diagnosed megalopolis is the mid-Atlantic one in the United States. By a broad areal definition only 6.0% of its territory was built up in 1960, 12% by a more restricted definition, and only 21% of the territory within its component Standard Metropolitan Statistical Areas (Regional Plan Association [21]). These statistics will be confirmed by anyone who has flown over the area and seen the vast open spaces within which the urbanized areas huddle. If there is a propensity for growth by the smaller units of megalopolis, there is no lack of space for that of growth.

I have argued in this paper that it is a mistake to emphasize the variation in urban costs with urban size and to neglect variations in productivity. Similarly, it is misleading to consider only size, which is a measure of immediate opportunities, while neglecting the broader context of opportunities in other cities. Big and small must be qualified in their setting: where it may be quite good to be smaller in a dense setting but it may be necessary to be quite big in an isolated one. Policies of small and far, which are not uncommon, perhaps should be small and near, and big and far.

Problems in the Interpretation of the Curves

It would be well to end this discussion with an upbeat note, but it must be concluded with further cavils of interpretation and definition. Curves such as those in Figure 1 will vary from city to city, from society to society, and within a society with its evolution through time. Even after the improvement of the contextual variable of income potential, the population variable accounts for only part of the variation in local incomes. Further, while the empirical data shows substantial consistency across the arrays of cities of particular nations, it is based primarily on cross-sectional data. One might ask what is the meaning for a city of 20,000 of curves extending into the tens of millions; or, conversely, what is the meaning for a great metropolis like New York or Tokyo of the range of the curve at the 20,000 level.

It must be granted that the relation of size to time is far from clear in these arguments. Surely it is not the same to grow tenfold in a decade as in a century, and movements along the population axis in the Figure would take place in real time and in particular circumstances. One might ask, further, whether paths or expansion coincide with paths of retrenchment, whether a city grown too big can regain paradise by shedding its excess population. In brief, the analysis admittedly suffers from the common limitations of the application of static theory to a dynamic process. Most particularly, there is no suggestion here that a static general equilibrium approach can serve to model a nation's system of cities because the movements of people, capital, ideas and institutional forms are slow and evolutionary, and while they change, other things are changing. Issues such as these have more to do with the dynamics of fluids than with the mechanics of solids.

BIBLIOGRAPHY

(1) Alonso, W. "Equity and Its Relation to Efficiency in Urbanization," in J. F. Kain (Ed.), Essays in Regional Economics. Cambridge: Harvard University Press, forthcoming.

(2) Alonso, W. "Industrial Location and Regional Policy," Working Paper No. 74, Center for Planning and Development Research, University of California, Berkeley, 1968.

(3) Alonso, W. "Urban and Regional Imbalances," Economic Development and Cultural Change, Vol. 17, No. 1 (October 1968). pp. 1-14.

(4) Alonso, W. and M. Fajans. "Cost of Living and Income by Urban Size," Working Paper No. 128, Center for Planning and Development Research, University of California, Berkeley, July 1970.

(5) Alonso, W. and E. Medrich. "Spontaneous Growth Centers in Twentieth Century American Urbanization," in N. Hansen (Ed.), Growth Centers and Regional Development (tentative title). New York: The Free Press, forthcoming.

(6) Berry, B. J. L. and E. Neils. "Location, Size, and Shape of Cities as Influenced by Environmental Factors: The Urban Environment Writ Large," in H. Perloff (Ed.), The Quality of the Urban Environment. Baltimore: The Johns Hopkins Press, 1969.

(7) Burns, L. S. "The Intra-Urban Income Distribution," paper presented at the Annual Conference, British Section of the Regional Science Association, London, 1968.

(8) Cameron, G. C. "Growth Areas, Growth Centres, and Regional Conversion," Scottish Journal of Political Economy, Vol. 21 (February 1970) pp. 19-38.

(9) Douglas, R. "Selected Indices of Industrial Characteristics for U.S. Standard Metropolitan Statistical Areas, 1968," Discussion Paper No. 20, Regional Science Research Institute, Philadelphia, 1967.

(10) Fuchs, V. R. "Differentials in Hourly Earnings by Regional and City Size, 1959," Occasional Paper 101, National Bureau of Economic Research, New York, 1967.

(11) Gallup, G. "Typical U.S. Family of Four Needs $120 per Week to Make Ends Meet," American Institute of Public Opinion, Princeton, New Jersey, 1970.

(12) Hirsch, W. Z. "The Supply of Urban Public Services," in H. Perloff and L. Wingo (Eds.), Issues in Urban Economics. Baltimore: The Johns Hopkins Press, 1968.

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

(14) Kain, J. F. "Urban Form and the Cost of Urban Services," Discussion Paper No. 6, Program on Regional and Urban Economics, Harvard University, 1966.

(15) Khorev, B. "What Kind of City is Needed," Literaturnaya gazeta, No. 14, 1969, as reported in The Current Digest of the Soviet Press, Vol. 21, No. 14 (April 23, 1969), pp. 14, 27.

(16) Mera, K. "On the Concentration of Urbanization and Economic Efficiency," Economics Department Working Paper No. 74, International Bank for Reconstruction and Development, Washington, D.C., March 1970, Monograph.

(17) Morse, R. et al. "Costs of Urban Infrastructure as Related to City Size in Developing Countries: India Case Study," Stanford Research Institute, Menlo Park, California, 1968.

(18) Ornati, O. A. "Poverty in Cities," in H. Perloff and L. Wingo (Eds.), Issues in Urban Economics. Baltimore: The Johns Hopkins Press, 1968.

(19) Perevedentsev, V. "Cities and Years," Literaturnaya gazeta, No. 9, 1969, as reported in The Current Digest of the Soviet Press, Vol. 21, No. 9 (March 19, 1969).

(20) Pred, A. The Spatial Dynamics of U.S. Urban and Industrial Growth, 1800-1914. Cambridge: M.I.T. Press, 1966.

(21) Regional Plan Association. "The Atlantic Urban Seaboard: Development Issues and Strategies," Regional Plan News No. 90, New York: Regional Plan Association, September 1969.

(22) Schmandt, H. J. and G. R. Stephens. "Local Government Expenditure Patterns in the U.S.," Land Economics, November 1963, Vol. 39.

(23) Schnore, L. "Some Correlates of Urban Size: A Replication," American Journal of Sociology, September 1963, Vol. 69, pp. 185-193.

(24) Thompson, W. R. "Internal and External Factors in the Development of Urban Economies," in H. Perloff and L. Wingo (Eds.), Issues in Urban Economics. Baltimore: The Johns Hopkins Press, 1968.

(25) Thompson, W. R. A Preface to Urban Economics. Baltimore: The Johns Hopkins Press, 1965.

(26) Vernon, R. Metropolis 1985. New York: Doubleday Anchor, 1963.

(27) Williamson, J. "Regional Inequality and the Process of National Development," Economic Development and Cultural Change, July 1965, Vol. 13, Part 2.

(28) Zadeh, L. A. "Toward a Theory of Fuzzy Systems," Report No. ERL-69-2, Electronics Research Laboratory, University of California, Berkeley, 1969.