Derivation of Some Social-Demographic Regularities from the Theory of Time-Minimization

G. EDWARD STEPHAN, *Western Washington University*

ABSTRACT

Earlier work on the size–density hypothesis led to a theory of time-minimization from which the size–density relation could be derived. Subsequently, time-minimization theory was employed to derive expected relations between population and area for cities and urbanized areas, expectations which were empirically confirmed. The present paper derives three well-known and empirically supported relationships from the time-minimization assumption: the gravity model of interaction, the intra-urban density function, and the rank–size rule for cities.

The size–density hypothesis was developed empirically from a study of the development of county boundaries in the United States from 1790 through 1970. Stated simply, the finding was that county size was inversely related to regional density throughout the period, with the relation weakening somewhat after the introduction of the automobile (Stephan:a). Subsequent studies, plus a few earlier ones (Callan and Stephan; Haggett; Myers and Stephan; Skinner; Stephan, b; Stephan and Tedrow, a; Stephan and Wright; Webb) have provided support for the hypothesis, both internationally and cross-culturally. In the most extensive of these studies (Stephan,b), involving the 1764 political divisions of 98 modern nations, the overall regression slope between logarithms of size and density was found to be negative two-thirds. This general relationship was later theoretically derived (Stephan,c) from the assumption that "social structures evolve in such a way as to minimize the societal time expended in their operation." Minimizing the sum of two opposing time-costs (travel-time and maintenance-time) led to the expectation (previously confirmed, of course) that

$$A = KD^{-2/3} \qquad (1)$$

where A is area, D is density, and K is a constant estimated from the data. A subsequent study (Massey and Stephan) showed that the size–density exponent for counties in Great Britain—which had proved a somewhat deviant case in the earlier international study, with a value of only -0.09—approached the theoretically expected value as one proceeded backwards through time to the census of 1801; the expected value was also very nearly approached by the reorganization of English local government which took place just a few years ago. Thus, time-minimization seemed sufficient to account for the size–density relation, as a general rule and in the case of a major apparent exception to that rule.

The same assumption proved fruitful in another sphere of inquiry, namely, the relation between urban population and land area (Stephan and Tedrow,b). We originally set out to account for Stewart and Warntz's empirical finding of a positive 0.75 slope between logarithms of population and area for cities in the United States in 1940 and in Great Britain in 1951 (Best et al. reported findings on cities in Great Britain from which one can compute a slope of 0.80). Using three time-cost terms (involving travel, competition for land, and competition for retail markets) we were able to derive two formulas under the assumption of time-minimization:

$$A_c = KP_c^{.77} \tag{2}$$

for central cities, and

$$A_u = KP_u^{.89} \tag{3}$$

for urbanized areas. Again, K was a constant estimated from the data. The previously published findings did not differ significantly from the expectation in equation (2). Each formula was supported by data from the 1950, 1960, and 1970 U.S. Censuses (the only ones in which the distinction of the two types of urban entities has been made), and the two types proved to be statistically distinguishable by this criterion.

The purpose of the present paper is to test still further the general applicability of the time-minimization assumption. My purpose is not to derive new hypotheses, nor even to test old ones, but rather to show that three fairly well-known and empirically supported findings can be derived from the assumption stated earlier, that social structures evolve in such a way as to minimize the societal time expended in their operation. The three relations to be derived are known as the gravity model, the urban density gradient, and the rank–size distribution. Owing to space limitations, these empirical relationships will be only very briefly described, relying primarily on reference to review literature rather than detailed citations of original sources. After this brief introduction, the relationships will each be derived from the assumption of time-minimization.

Empirical Formulas

The *gravity model* (or interactance hypothesis) says that interaction between two places will be a direct function of the product of their populations and an inverse function of the distance between them. The traditional formulation is

$$I_{ij} = kP_iP_j/D_{ij} \qquad (4)$$

where I_{ij} is the amount of interaction (e.g., migration, travel, communication, freight traffic) between places i and j, P_i and P_j are the respective populations, and D_{ij} is the distance. K is a constant of proportionality, estimated empirically from regression analyses of other factors. The earliest explicit statement of the model was by Carey. It was given formal structure, extensive empirical support and popularization by Stewart and Zipf, each following direct analogies from the gravity formula in physics. Carrothers lists 83 bibliographic references to works which either suggest modifications of the basic formula (e.g., weighting factors or exponents for any of the terms) or provide empirical tests of its utility. More recent reviews (e.g., Olsson) indicate how enormous the body of empirical work on the gravity model has become.

The *urban density gradient* was given formal expression by Clark, although the discussion of population decline around cities dates back at least to the work of von Thünen in 1826 (see Edmonston; Haggett). The expression was derived empirically; it is usually stated in the form

$$D_x = D_oe^{-bX} \qquad (5)$$

where D_x is the residential density at the distance X, D_o is the "central density" (simply the density intercept, usually estimated through measurements of densities for outlying areas), e is the base of the natural logarithms, and b is the rate of change in density per unit distance. On the basis of a study of 36 cases, Clark argued that the negative exponential density function appeared to hold "for all times and places studies, from 1801 to the present day, and from Los Angeles to Budapest" (475). Subsequent empirical work, summarized by Berry et al., has shown very widespread support for Clark's hypothesis, historically and cross-culturally. They conclude with a reprise of Clark's assessment: "Regardless of time or place, the expression . . . provides a statistically significant fit to the distribution of population densities within cities" (403).

The *rank–size rule* states that, when cities are rank-ordered according to size, from the largest to the smallest, the product rank-times-size approximates a constant equal to the size of the largest city:

$$Rs_r = K = s_1 \tag{6}$$

where R is rank, s_r is the size of the rth-ranked city, and K is a constant equal to s_1, the size of the largest city. The relation was first noted by Auerbach for the 94 largest cities in the German Census of 1910. Lotka reformulated the rule, with an exponent for R which need not be unity as in Auerbach's formulation. Zipf popularized the rank–size rule, providing numerous empirical studies and stimulating a good many more. The rule has been found to hold for cities in the United States, Canada, most European nations, Japan, Malaya, India, and even to cities of the world taken as a whole (Gossman). An extensive reivew of attempts to derive the rule theoretically is given by Richardson.

Derivations

GRAVITY MODEL

We begin with an individual located at place i who for whatever reason is unable to obtain what he needs at that location and is hence compelled to consider traveling to, or to and from some other place j. We assume that the probability of going from place i to place j will be inversely proportional, first, to the time it takes to get from i to j (and back if need be), which we will symbolize T_{ij}, and second, to the time it takes to locate what he needs at j when he arrives, which we will symbolize as T_j. We can express this assumption as

$$p_{ij} = c/(T_{ij}T_j) \tag{7}$$

where p_{ij} is the probability of going from i to j, and c is the constant of proportionality.

We can specify the equation somewhat by noting that T_{ij} is simply the distance D_{ij} divided by the average velocity of the means of transportation, v. Further, we know empirically (e.g., Berry) that the larger a place is the more likely one is to find the quantity and variety of goods and services one might need (except, of course, for the need to escape from large populations—see Catton, for just such a study, travel to national parks). With T_j inversely proportional to P_j, we can rewrite the above expression as

$$p_{ij} = vcP_j/(sD_{ij}) \tag{8}$$

where s is the constant of proportionality between T_j and P_j.

If we now let the single constant k stand for the constants vc/s, and if we multiply the probability of going from i to j by the "population at

risk," P_i, we obtain an expression for I_{ij}, the amount of interaction,

$$I_{ij} = p_{ij}P_i = kP_iP_j/D_{ij} \tag{9}$$

the traditional statement of the interactance hypothesis or gravity model.

The simple model can, of course, be made much more complex as noted above (see Carrothers). The present derivation, in fact, suggests one additional modification: specification of the constant k in terms of v, c and s. As transportation improves so that v increases, as c (presumptively the relative attractiveness of j over i, independent of size) increases, and as s (perhaps the congestion or competition generated by P_j) decreases, interaction increases. The main point of the present exercise, however, is simply to demonstrate that the gravity model, in its elementary form, can be derived from the time-minimization assumption.

URBAN DENSITY GRADIENT

Assume[1] that the unit-cost of land occupancy declines exponentially with distance from the center of the city:

$$U_x = U_oe^{-aX} \tag{10}$$

where U_x is the cost per unit of land at the distance X, U_o is the unit-cost at the center of the city, and a is the rate of change in unit-cost per unit of distance from the center.

Total land-occupancy cost at the distance X will be the unit-cost times the area occupied at X, i.e., the product A_xU_x (where A_x is area occupied at X). This cost is usually expressed in monetary units; but, since money is received in units of money per unit of time (e.g., dollars per hour of labor time, investment time, rental time, etc.), we can express the time-cost of land occupancy as hA_xU_x (where h is the inverse of income-per-unit-time, e.g., hours per dollar) to obtain

$$T_{rx} = hA_xU_oe^{-bX} \tag{11}$$

where b is the parameter a adjusted to reflect declines in hourly (as opposed to monetary) costs per unit distance from the center and T_{rx} is the time-cost of residing at the distance X.

The travel-time expended by the population residing at the distance X, in moving to and from the center, will be the distance, times the population P_x, divided by v, the average velocity of the means of transportation employed. This aggregate travel-time can be expressed as

$$T_{tx} = XP_x/v \tag{12}$$

which, combined with expression (11) gives

$$T_x = T_{rx} + T_{tx} = hA_xU_oe^{-bX} + XP_x/v \tag{13}$$

as the expression for total time-expenditure required at the distance X.

Differentiating this expression with respect to X yields

$$dT_x/dX = P_x/v - hA_xU_obe^{-bX} \tag{14}$$

and this, set equal to zero and solved for $D_x = (P_x/A_x)$ gives us

$$D_x = vhbU_oe^{-bX} \tag{15}$$

as the relation between density and distance when total time is minimized (the formal condition for a minimum, that the second derivative be greater than zero, is satisfied since h, A_x, U_o and b itself are non-negative). Finally, if we let the constant D_o stand for the product of the constants v, h, b and U_o, we have the formula

$$D_x = D_oe^{-bX} \tag{16}$$

i.e., the traditional formula for the urban density function.

There are several points to note in connection with this derivation. First the *central density* D_o has been given an explicit definition as the product of the factors v, h, b and U_o. Taken separately, an increase in any of these factors should lead to an increase in D_o. But the factors can seldom be taken separately. Thus, increases in average velocity have tended to be associated with decreases in the magnitude of b. Further, decreases in h (i.e., an increase in hourly income) should be related to increases in U_o (since more money would be available with which to bid for land) and v (since more money would, presumably, lead to more efficient means of transportation). The relation is complex and its analysis lies beyond the scope of this paper (but see Edmonston for a detailed discussion). The definition of D_o given here is an improvement over that usually given (the simple inward extrapolation of outlying densities) in that it specifies these factors and does not pretend to represent a "central density." (We know, in fact, that the central residential density tends to be much lower than D_o as usually constructed from regression analyses; see Haggett; Newling.)

Second, the *density gradient*, b, has been given some definition, beyond the simply empirical role it plays as an estimate of the density gradient. Presumptively, in our derivation, it mirrors (through the time-money equivalency in h) the decline in unit-costs of land given by parameter a. Technically, the symbols U_o and a describe land costs for all possible uses: commercial, industrial and residential. Either because potential resi-

dential users lack the means to compete in the central land market, or because they are prevented from doing so through zoning, it frequently happens that costs for *residential* land *decline* near the center of the city (i.e., there is less of a residential land market there; see Berry et al.). In this case the function describing land costs for residential use would not be negative exponential as in equation (11). It would rise from a central low, then peak, then decline (as with lognormal, chi-square, or certain Weibull functions for example). Whatever the function, it could still be brought through the same type of derivation as the negative exponential has been here, with high densities paying the cost of expensive land as a result of time minimization.

RANK–SIZE RULE

Whatever its specific content, human interaction takes time. As the size, s, of an interacting group increases, the number of possible interactions between each member and all others increases by a factor of $s(s-1)$ or, as s becomes large, by a factor of s-squared. Given that each individual has only a finite amount of time available for interaction, we expect that as groups increase in size the *potential* for total interaction becomes more and more impossible to realize. We accommodate through role-specific interactions, regulations and authority relations, market interactions, and a host of similar institutional structures. But all of these structures themselves involve time-expenditure for their operation, so the factor s^2 can still be said to be a measure of time-expenditure for the group. If we had only to minimize such time-expenditure, the solution would be a set of N uniformly small cities, with N times s representing the total population.

Still, there are advantages in having at least some large cities in a society. Large cities can provide a focal point for the assembly of raw materials from elsewhere, easy access to large labor markets for manufacturing and large retail markets for the distribution of specialized goods and services, and economies of agglomeration resulting from the clustering of many related activities in one place. In effect these advantages represent time-*savings* associated with increased size. If we had only this factor to consider, time-minimization could be achieved by having but one city which contained the entire population. We thus have two contradictory solutions to the problem of time-minimization: numerous equally small cities or one enormous city.

A compromise might be suggested, namely, a fairly small number of fairly large cities, all of some average or optimum size. But this solution denies society both the time-savings of very large cities and the time-savings of numerous small cities. A solution which comes as close as possible to providing both kinds of time-savings is one in which there is considerable variation in city size, but in which the likelihood of finding

cities of a specific size is inversely proportional to their interaction-time-cost. If we let $p(s)$ be the probability of finding a city of size s, we are saying that $p(s)$ is to be inversely proportional to s^2 (measure of interaction-time-cost). With c as the constant of proportionality, we have

$$p(s) = c/s^2 \tag{17}$$

We assume a distribution of city sizes, from s_1, the largest, down to s_n, the smallest, with s_r ($r = 1,2, \ldots,n$) being the rth-ranked city in this distribution. The cumulative probability distribution can be obtained from the above expression through integration, with theoretical limits of s_n to infinity; thus

$$P(s) = \int_{s_n}^{\infty} c/s^2 ds$$

$$= c/s_n \tag{18}$$

which, over its total range, must equal 1.0, yielding $c = s_n$. Substituting this value, we can determine the probability of finding a city equal to or less than size s^r as follows

$$P(s \leq s_r) = \int_{s_n}^{s_r} s_n/s^2 ds$$

$$= 1 - s_n/s_r \tag{19}$$

The probability of finding a city *greater than or equal to* s_r is simply one minus this expression, or s_n/s_r. This probability, times the number of cities N, gives the rank R so that

$$R = NP(s \geq s_r) = Ns_n/s_r \tag{20}$$

or, multiplying through by s_r and letting K stand for the product of the given constants N and s_n, we have

$$Rs_r = Ns_n = K = s_1 \tag{21}$$

the traditional forms of the rank-size rule.

This expression can be further complicated if we assume that the two economies which we tried to compromise are, for whatever technological reasons, not equally important. If we let d stand for the importance of time-savings due to *decreasing* interaction (reduction of s^2) and if we let i stand for the importance of time-savings realized through *increasing* interaction (the economies of agglomeration and so on mentioned above), we

could begin our derivation, not with s individuals interacting with s other individuals (s^2), but rather with s individuals interacting with $s^{d/i}$ individuals. Then, minimizing the interaction term $s^{1+(d/i)}$ as before, we would obtain a final result of the form

$$R^{i/d}s_r = K \qquad (22)$$

where the exponent i/d could vary from zero to infinity. In the former case, if d were the only factor of importance, we would obtain the solution with N equally small cities; in the latter, if i were the only factor of importance, we would obtain the solution with one enormous city. These results, and those with intermediate values of i/d, all result from the assumption of time-minimization. We thus complete the set of derivations to be attempted here.

Conclusions

My point in attempting these derivations has not been simply to derive the previously known formulas. A large and varied set of such derivations already exists, and there would be little point in adding yet another. Rather, my purpose has been to show that five *independent* empirical generalizations:

1. size–density hypothesis
2. urban area-population relationships
3. gravity model of interaction
4. urban density function
5. rank–size distribution of cities

can all be derived from the *same* theoretical assumption, namely, that social structures evolve in such a way as to minimize the societal time expended in their operation.

A previous effort to derive a number of disparate findings from a common theoretical assumption deserves comment here. George Zipf argued that a number of patterns in human behavior could be derived from what he called the *principle of least effort*. The conception has proved useful because effort can stand for many mixed forms of cost (distance, time, money, energy). As Abler et al. put it, "The most direct path in a distance sense between a chair behind a desk and the hallway might be a straight line through the desk and the wall, yet the least effort (least time and cost) path goes around the desk and through the door" (253). Since we usually cannot maximize or minimize several variables simultaneously (the problem of the greatest happiness of the greatest number has no unique solution), we aggregate all our cost variables into something called effort and reason

that that is what we minimize. The problem with Zipf's formulation, I believe, is that the concept of effort is too vague to serve as a basis for deductive theory. A solution, the one we have employed here and earlier, is to pick one cost-factor and attempt to express the others in its dimension. One can usually specify equivalencies between cost-factors like space, time, money, manpower or horsepower, at least in a general way. The question then is, Why pick time as the fundamental dimension?

If the time available to human beings were infinite, all other cost factors would be irrelevant. If time were infinite, any distance could be traversed however slowly, any amount of money could be accumulated sooner or later, monumental tasks could be accomplished using very rudimentary technology. It is the fact that time is limited which gives the other cost factors, and various means of reducing them, their significance. In this sense the time dimension is fundamental. The time dimension is also clearly panhistoric and cross-cultural; any society, viewed in this way, is a finite budget of man-hours expended in characteristic subsistence and non-subsistence activities under given environmental and technological conditions. Thurstone, I believe, once said something to the effect that whatever exists exists in some quantity; in a parallel sense, whatever human beings do they do through some expenditure of time. And if time is the ultimately limited resource, then it seems sensible to consider the time dimension as fundamental to our theorizing.

I believe many of the empirically supported findings of sociology, at levels of analysis other than the social-demographic (e.g., from the literature of formal organization research) could be theoretically derived from a sumptions of time minimization. Even findings at the social psychological level of interpersonal behavior might be amenable to such analysis (e.g., the proximity-similarity-attraction findings over the last few decades: it takes less time to meet people nearby, then, after longer periods of search and discovery, it takes less time to interact among those who are similar than among those who are not).

In the end, as Comte argued in establishing the field, the purpose of theorizing is to reduce as many empirical findings as possible to as few theoretical assumptions as possible. I believe the assumption of time-minimization deserves serious consideration from theoretical sociologists, at least to the degree that movement-, punishment-, and cost minimization have received attention from theoretical geographers, psychologists, and economists.

Note

1. The assumption develops from the following rationale. Presumably, bid-rents for land—the amount of money a set of potential users are willing to pay for such use—declines *linearly* as a function of distance from the central place, as a direct consequence of the time-cost of travel to the central place. But the importance of such time-costs varies from one *type* of user to another. A central location is more important to a commercial than a residential user of

land, so the potential commercial user will outbid the potential residential user—i.e., the line relating distance to bid-rent will be higher at the center and have a steeper decline for the commercial user than for the residential user. A *family* of such lines is said to have as its *envelope* that curve to which each of the lines is tangent. (See Botyanskii for a treatment of envelopes which does not require work in differential equations.) Though the family of lines suggested here should in general have a downward sloping envelope with a finite y-intercept and an asymptotic approach to the x-axis, it is not necessary that the curve be negative exponential; but such a curve does at least satisfy the conditions just mentioned, and it probably describes empirical conditions as well as any others (see Abler et al. 357–65, for illustrations and for the relation between such a curve and concentric zone theory).

References

Abler, Ronald, John S. Adams, and Peter Gould. 1971. *Spatial Organization: The Geographer's View of the World*. Englewood Cliffs: Prentice-Hall.

Auerbach, F. 1913. "Das Gesetz der Bevolkerungskonzentration." *Petermanns Mitteilungen* 59(February):74–6.

Berry, Brian J. L. 1967. *Geography of Market Centers and Retail Distribution*. Englewood Cliffs: Prentice-Hall.

Berry, B. J. L., J. W. Simmons, and R. J. Tennant. 1963. "Urban Population Densities: Structure and Change." *Geographical Review* 53(July):389–405.

Best, R. H., A. R. Jones, and A. W. Rogers. 1974. "The Density-Size Rule." *Urban Studies* 11(June):201–08.

Boltyanskii, V. G. 1964. *Envelopes*. Oxford: Pergamon (originally published as *Ogibayushchaya*, Moscow, 1961; translated and edited by Robert B. Brown).

Callan, J. S., and G. E. Stephan. 1975. "Siwai Line-Villages: Thiessen Polygons and the Size-Density Hypothesis." Solomon Island Studies in Human Biogeography. Occasional paper number 4, Field Museum of Natural History, Chicago.

Carey, H. C. 1858. *Principles of Social Science*. Philadelphia: Lippincott.

Carrothers, G. A. P. 1956. "An Historical Review of the Gravity and Potential Concepts of Human Interaction." *Journal of the American Institute of Planners* 22(Spring):94–102.

Catton, William R., Jr. 1966. *From Animistic to Naturalistic Sociology*. New York: McGraw-Hill.

Clark, C. 1951. "Urban Population Densities." *Journal of the Royal Statistical Society*, Series A, 114:490–96.

Edmonston, Barry. 1975. *Population Distribution in American Cities*. Lexington, Mass.: Heath.

Gossman, Charles S. 1966. "The Rank-Size Rule and Mathematical Models for the Size Distribution of American Cities." Unpublished M.A. thesis, University of Washington.

Haggett, Peter. 1965, *Locational Analysis in Human Geography*. London: Arnold.

Lotka, Alfred J. 1924. *Elements of Physical Biology*. Baltimore: Williams and Wilkins (reissued as *Elements of Mathematical Biology*, by Dover, 1956).

Massey, D. S., and G. E. Stephan. 1977. "The Size-Density Hypothesis in Great Britain: Analysis of a Deviant Case." *Demography* 14(August):351–61.

Myers, D. E., and G. E. Stephan. 1974. "Tribal Territories of the California Indians: A Test of the Size-Density Hypothesis." *Anthropology UCLA* 6(Fall):59–66.

Newling, B. 1969. "The Spatial Variation of Urban Population Densities." *Geographical Review* 59(April):242–52.

Olsson, Gunnar. 1965. *Distance and Human Interaction: A Review and Bibliography*. Philadelphia: Regional Science Research Institute.

Richardson, H. W. 1973. "Theory of the Distribution of City Sizes: Review and Prospect ." *Regional Studies* 7(September):239–51.

Skinner, G. W. 1964. "Marketing and Social Structure in Rural China." *Journal of Asian Studies* 24(November):3–43.
Stephan, G. E. a:1971. "Variation in County Size: A Theory of Segmental Growth," *American Sociological Review* 36(June):451–61.
_____ b:1972. "International Tests of the Size-Density Hypothesis." *American Sociological Review* 37(June):365–68.
_____ c:1977. "Territorial Division: The Least-Time Constraint Behind the Formation of Sub-National Boundaries." *Science* 196(29 April):523–24.
Stephan, G. E., and L. M. Tedrow. a:1974. "Tribal Territories in Africa: A Cross-Cultural Test of the Size-Density Hypothesis." *Pacific Sociological Review* 17(July): 365–69.
_____ b:1977. "A Theory of Time-Minimization: The Relationship Between Urban Area and Population." *Pacific Sociological Review* 20(January):105–12.
Stephan, G. E. and S. M. Wright. 1973. "Indian Tribal Territories in the Pacific Northwest: A Cross-Cultural Test of the Size-Density Hypothesis." *Annals of Regional Science* 7(June):113–23.
Stewart, J. Q. a:1941. "An Inverse Distance Variation for Certain Social Influences." *Science* 93(January):89–90.
_____ b:1948. "Demographic Gravitation: Evidence and Applications." *Sociometry* 11(Spring):31–58.
Stewart, J. Q., and W. Warntz. 1958. "Physics of Population Distribution." *Journal of Regional Science* 1(Summer):90–123.
Webb, S. 1974. "Segmental Urban Growth: Some Cross-National Evidence." *Sociology and Social Research* 58(July):387–91.
Zipf, George K. a:1946. "The P_1P_2/D Hypothesis: On the Intercity Movement of Persons." *American Sociological Review* 11(December):677–86.
_____ b:1949. *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley.