ON PROBLEMS OF MEASURING THE DISTRIBUTION OF POPULATION IN AN URBAN AREA*

Martin Taitel,** University of Chicago

The fact of human life with which we are at present concerned is this: A very large proportion of the people globulates. Human beings have not spread equally over the earth's surface (total or land). Nor have they spread unevenly with gradually and smoothly rising and falling densities. Rather they form globules or near globules of population at selected locations. These are loosely and variously referred to as cities, standard metropolitan areas, urban places, urbanized areas, and metropolises.

Quite obviously the phenomenon of population globules presents many complex facets for analysis. Variables are many--both those within globules and those as between globules. Within an individual globule, there is variation in racial composition, income, wealth, land value, density, and so on. Comparison of globules shows variation in size, income, wealth, land value, transport facilities, average density for finite areas, industrial composition, and so on. Relations--including those between the "within globules" and those between the "between globules" variables--do not appear to be simple or easy to measure.

This paper is limited to consideration of the density and distance variables of individual globules. Conceptualization of location as a point in a plane and of density as a count of persons per unit area (finite or infinitesimal) of the plane defines the variables under consideration. For those variables, this paper is concerned with methods of measuring the relation between density and location within an urban area --with methods of measuring the shape of an individual population globule.

It is convenient at this point to set forth some of the notation to be used. This is shown in Table I. It should be noted that provision is made for the distinction between "density" and "probability."

According to Winsborough (Unpublished, 1960), there is a considerable body of data of varying

TABLE I

SELECTED SYMBOLS USED

Quantity	Origin or pole: O(True)	O'(False)
Probability	P	P
Density for persons:		
At a point	D	D
At the origin or pole	D_0	D_0'
Density for an area	δ	δ
Polar Coordinates:		
Radius	ρ	ρ'
Direction	θ	θ'
Direction OO'	θ	$*\phi$
Distance OO'	0	*C
Rectangular Coordinates:		
Abcissa	x	X
Ordinate	y	Y
X - x	0	$*C_x$
Y - y	0	$*C_y$
Parameter for exponential function	μ	μ
Size of total universe	N	N
Size of sample	n	n
Area	A	A
Frequency:		
Theoretical	f	f
Observed	f'	f'

*Parameters of location on the plane approximating the earth's surface.

quality bearing on density and location within urban areas. Yet Duncan (1957) comments: "Surprisingly little systematic study has been devoted to the pattern of variation of population density from one part of the city to another." This applies to both empirical measurements of and theorizing about the shape of population globules. Since Duncan wrote in 1957, there has been some additional published work by Clark (1958) and Stewart and Warntz (1958) and within my knowledge some still unpublished work by Muth and Winsborough.

In this limited amount of study, there has developed an hypothesis expressing density as a bivariate exponential function of location. According to Duncan (1957) it was first suggested by Stewart (1947) and later by Clark (1951). In our notation, this is

---

*Presented before the session, Metropolitan Areas and Urban Growth--II, at the 120th annual meeting of the American Statistical Association, Stanford University, August 25, 1960.

**I wish to express my appreciation to Professor Philip M. Hauser, Director, Population Research and Training Center, University of Chicago, for making available to me the facilities of the Center in the preparation of this paper. I wish also to express my appreciation to Mr. Halliman H. Winsborough, Research Assistant at the Center, for time spent in discussion of many aspects of urban densities and of the literature relating to them and for making work sheets available to me. It seems appropriate to state also that my efforts represent a labor of love and not of duty.

I wish, in addition, to state that this paper has benefited much from the comments on an earlier draft made by Professors Philip M. Hauser, William H. Kruskal and Richard Muth, all of the University of Chicago.

$$D = \frac{1}{2\pi}\left(\frac{2}{\mu}\right)^2 e^{-\frac{2}{\mu}\rho}, \begin{pmatrix} 0 \leq \rho \leq \infty \\ 0 \leq \theta \leq 2\pi \end{pmatrix} \quad (1)$$

the variables of course being D, ρ, and θ. The listener is cautioned to remember that ρ and θ are polar coordinates and that (1) is not a probability function.*

Note that in formulating the hypothesis in this way ρ and θ and, consequently, D are specified as continuous variables. Further, D is a relative density at a point, not an absolute density. Absolute densities may be conceived of as given by ND. However, since N can in fact only be finite, consistency with fact must be obtained through a conceptualization of persons as divisible into infinitesimal particles.

The hypothesis (1) can be cast in terms of probability. If this is done, ρ and θ become rectangular coordinates and

$$P = \frac{1}{2\pi}\left(\frac{2}{\mu}\right)^2 \rho\, e^{-\frac{2}{\mu}\rho}, \begin{pmatrix} 0 \leq \rho \leq \infty \\ 0 \leq \theta \leq 2\pi \end{pmatrix} \quad (2)$$

Integrating with respect to θ gives a Gamma distribution as the univariate distribution for ρ.**

Properties of (1) and (2) are as follows:

1. μ is the only parameter (since the total volume and the origin are assumed known).
2. μ is the arithmetic mean distance from the center, 0, or the first moment of ρ.
3. $\mu/2$ is the modal distance from the center, 0, or the modal ρ.
4. The maximum density (not probability) is

$$D_0 = \frac{1}{2\pi}\left(\frac{2}{\mu}\right)^2 \quad (3)$$

5. ρ and θ are statistically independent. Thus, distance from the center does not depend upon direction from the center and vice versa.***
6. If (1) is expressed in rectangular coordinates, then x and y are not statistically independent. This statistical dependence represents a desirable feature of the hypothesis.*

With (1) and (2) as a working hypothesis the next step is the measurement of parameters. The statement of the hypothesis assumes that the center or pole, 0, is known so that measurements are from that pole. Hence, let us proceed, in the first instance, on that basis. Let us assume, further, that sampling is of persons in the globule and is random with respect to them, i.e., we have a random sample of ρ and θ distributed as (2). And, finally, let us assume no errors of observation. Under these assumptions, the maximum liklihood estimate for μ is

$$M = \frac{\sum_{i=1}^{n} \rho_i}{n} \quad (4)$$

It is interesting to note that sampling-wise $4nM/\mu$ is distributed as Chi-square for samples of size 4n. Consequently, the F-distribution is applicable for testing differences between sample M's. Furthermore, the Chi-square goodness-of-fit test seems to be applicable to test the hypothesis itself.

These results are not, however, a solution to the problems of measurement which arise in practice.

For one thing, cities are not all circular in shape; there seems to be no simple general procedure for adapting the technique in cases of noncircularity.

Second, in practice, sampling can only be over a finite range for ρ. This, however, is not a serious drawback. Table II shows the

TABLE II

FRACTION OF POPULATION BEYOND RADIUS r AND RATIO OF μ TO μ_N

r/μ	Fraction of population beyond radius r	μ/μ_N*
0	1.0000	∞
1	.4060	1.8372
2	.09158	1.1923
3	.01735	1.04757
4	.003019	1.01089
5	.000499	1.002276
6	.000080	1.0004426
7	.000012	1.0000815
8	.000002	1.0000144
9	.0000003	1.0000025
10	.00000004	1.0000004

*Approximate correction factor for M_r to estimate μ.

Note: μ_N is the average distance for $0 \leq \rho \leq N$

*I am indebted to Professor William H. Kruskal for suggesting the difference between the density and probability function be called to the attention of the audience.

**It may be helpful to the audience to note that, had (1) been expressed in rectangular coordinates, the density and probability functions would be the same.

***This is also true for the normal distribution analogue, i.e.,

$$K e^{-\frac{\rho^2}{\sigma^2}} \text{ (in } \rho \text{ and } \theta)$$

*For the comparable zero-correlation bivariate normal distribution, x and y are, of course, statistically independent. This represents a very great disadvantage of the normal hypothesis.

approximate average error in M_N (the mean when sampling is only out to $\rho = N$) as an estimate of μ because it does not take into account the fact that sampling is limited to $0 \leq \rho \leq N$. If observations are over a radius five times the average distance, the error is only somewhat more than 2 parts in 1000 on the average. In practice, therefore, either observations can be obtained over a large enough range to obviate the need for adjustment or an approximation to the maximum liklihood solution as given by

$$\left[1 - \frac{2\left(\frac{N}{M}\right)^2}{e^{2\left(\frac{N}{M}\right)} - 1 - 2\frac{N}{M}}\right] M = \frac{\sum_{i=1}^{n} \rho_i}{n} \qquad (5)$$

can be used.

Third, in practical applications, not only relative densities, but also absolute densities, are sought. To obtain measurements of them either the urban area must be completely enumerated or else a supplementary estimating procedure must be used. This difficulty is not insurmountable.

Fourth, there arises the question of the validity of the random sampling assumption.* It is not easy to resolve this question, particularly since it involves the question of specifying the universe of which the complete enumeration of an urban area can be a random sample. There appears to be no forthright solution.

The fifth and last point in this discussion of the problems arising in measuring μ relates to the necessity for generalizing the measurement procedure to include measurement of the location of the center of the urban area (i.e., the origin or pole for the location coordinates). The practice has been to designate a point as the center of a city on the basis of an inspection of the data or of preconceived notions about where the center lies. For example, Blumenthal (1949) indicates the City Hall was used as the center for his data on Philadelphia. Clark (1951 and 1958), Muth (1960) and Winsborough (1960) use some point in the central business district. Reinhardt (1950) used the city hall as the center. The practice is, of course, subject to serious shortcomings as a method of measuring parameters. It should not be used particularly when more efficient procedures based upon the theory of probability are readily available.

It is convenient to demonstrate the availability of a valid procedure for measuring the parameters of location in terms of rectangular coordinates.**

*Professor O. D. Duncan, Population Research and Training Center, University of Chicago, raised the question at first sight of the author's preliminary work.

**The choice of coordinate systems depends, of course, upon the one used in obtaining the observations. Which to use in the observation process depends in turn upon considerations relating to errors of observation--not under consideration at the present time.

The maximum liklihood equations for C_x and C_y are

$$\sum_{i=1}^{n} \frac{X_i - C_x}{\sqrt{(X_i - C_x)^2 + (Y_i - C_y)^2}} = 0 \qquad (6)$$

and similarly for C_y. It is interesting to note that these are the equations which locate the point from which the average distance for points on a plane is a minimum.

I see no straightforward solution of (6) for the estimates C_x and C_y. But clearly a successive approximation procedure can be used. The desired quantities are weighted arithmetic means. Unweighted arithmetic means* can be used as first approximations. On the basis of them, approximate weights can be introduced, and so on. Intuitively I believe the efficiency of such estimates to be high.

Efficient estimation of the center of an urban area is, of course, of some importance. The effect of an error in locating the center is to increase the estimate of the average distance from the center, i.e., the estimate M of μ, and to reduce the estimate of the peak density by the square of the factor involved. Unfortunately, I have been unable to perform the integrations necessary to show the relation between C and the error in M. C is, of course, the distance between true and false centers. However, it is clear that an error in M cannot exceed C and must be substantially less. Further, the square root of the second moment about the false center to the second moment about the true center is given by

$$\sqrt{1 + \frac{2}{3}\left(\frac{C}{\mu}\right)^2} \quad . \qquad (7)$$

This may serve as a guide to the order of magnitude of the error in M when computed from a false center.

In practice, various investigators could hardly err by as much as μ, assuming, of course, some validity to the basic hypothesis. And perhaps they do have sufficient insight and intuition to avoid going off-center by significant distances. Nevertheless, it is surprising to discover they had not at least raised the question of where the center of a city is. It would seem that investigators would have attempted to determine whether the center of a city is a court house, a department store, a hotel, a church, a factory, a railroad station, etc. It would seem they would have tried to determine whether the center of a city moves from time to time. And, in addition, it would seem they would have attempted to determine whether cities differed with respect to the activities located at their centers.

The procedure for measurement which I have been discussing has not been applied, though it is certainly sound and worthy of consideration. The observations required are of horizontal distance on a rectangular grid or of horizontal distance and direction on a polar coordinate grid.

*The solutions for the bivariate normal surface.

Densities as such are not observed. Nor need they be computed in the estimation process and χ^2- goodness-of-fit testing.

An alternate procedure--and the one used by other investigators--is to take D, i.e., density, as an observable variable. For the present, let us assume that there is a value D attached to each person in the population and that this value of D is the one given by his location in the (ρ, θ) or (x,y) system of coordinates. In addition, we assume, as before, no errors of observation.

Let us first view our measurement problem on the assumption that the center is accurately known. If this assumption, in addition to those previously made, holds and the hypothesis (1) is true, then each and every observation must fall on the line

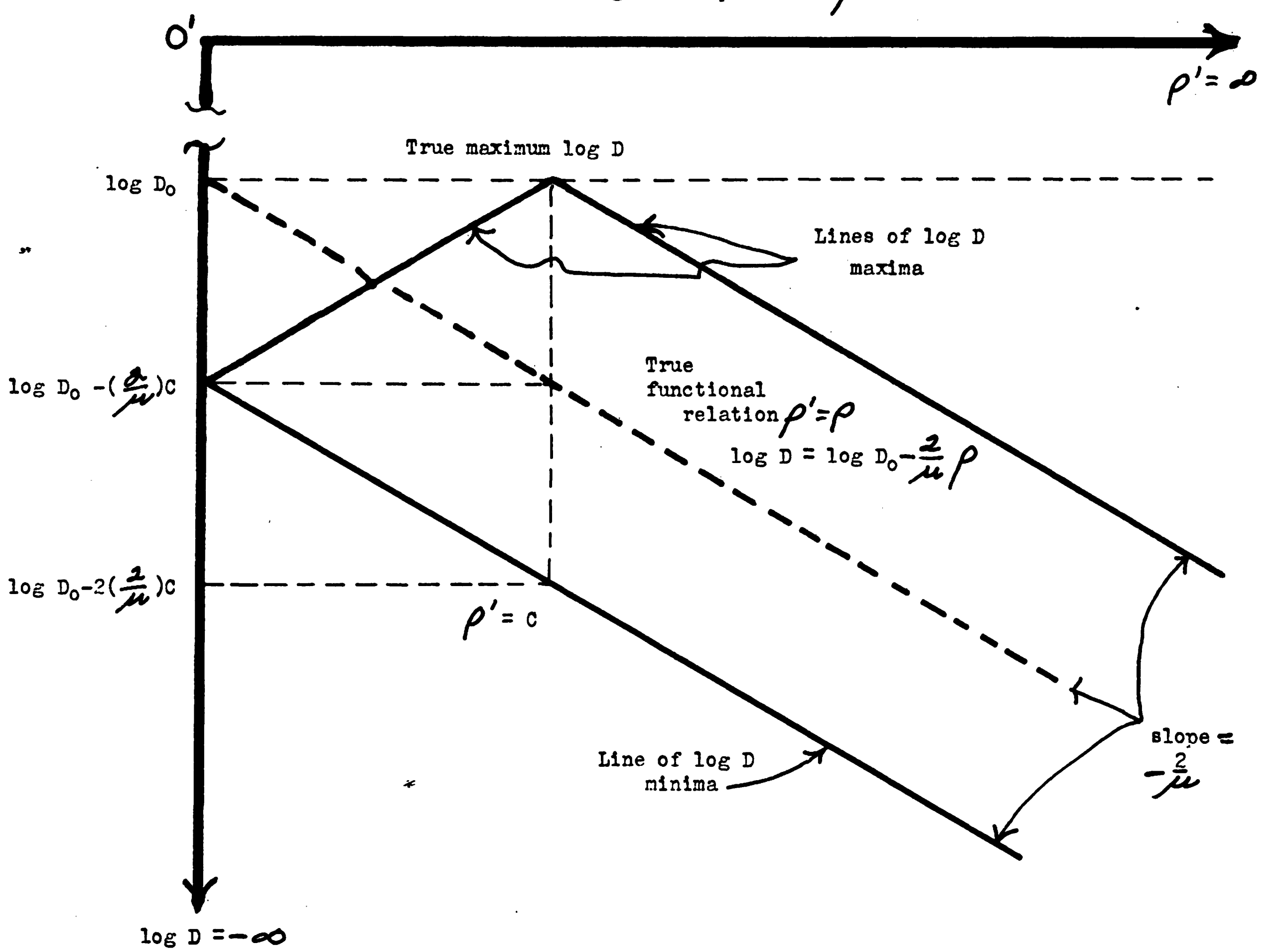
$$\log D = \log D_0 - \frac{2}{\mu}\rho \tag{8}$$

which is shown on the chart. If one observation is not on the line, then the hypothesis must be rejected. Furthermore, it may be noted that one observation is sufficient to determine μ, if relative densities are measured; and two non-identical observations are sufficient to determine μ and N, if absolute densities are measured. Thus, if a great deal is known, the parameters can be measured without regard to the character of the sampling.

To bring our view of the measurement problems somewhat more in accord with observational conditions let us assume the center of location is not known--still assuming, of course, that (1) is true and observations are perfectly accurate. Our variables then become log D and ρ', the latter being the distance from a "false" center. And the probability function for those variables is

Chart

Boundaries of Log D for Specified ρ'



$$P = \frac{e^{\log D}\left(\frac{\mu}{2}\right)^2\left(\log\frac{D_0}{D}\right)}{C\sqrt{1-\left[\frac{\left(\frac{\mu}{2}\right)^2\left(\log\frac{D_0}{D}\right)^2-\rho'^2-C^2}{2\rho' C}\right]^2}}, \quad 0 \leqq \rho' \leqq \infty \qquad (9)$$

$$\begin{aligned}\log D &\geqq \log D_0 - \frac{2}{\mu}C - \frac{2}{\mu}\rho' \\ &\leqq \log D_0 - \frac{2}{\mu}C + \frac{2}{\mu}\rho', \text{ for } 0 \leqq \rho' \leqq C \\ &\leqq \log D_0 + \frac{2}{\mu}C - \frac{2}{\mu}\rho', \text{ for } C \leqq \rho'\end{aligned}$$

where C is, of course, the distance between "true" and "false" centers. I regret that I have not had the opportunity to investigate this probability function. However, characteristics which I have been able to determine do shed some light on our measurement problem.

From a geometrical consideration of (1), the boundaries within which the universe of observations of log D and ρ' fall can be determined. These boundaries are shown on the chart and the equations for them have been written as part of (9). Actual observations, therefore, except very rarely by chance or unless deliberately selected to do so, would present a scatter. Neither the complete universe nor a sample (except extremely rarely) would appear as the functional relation (8) between density and distance. Furthermore, for the range $\rho' < C/2$, all observations would fall below the theoretical functional relation (8). Finally, it may be noted that, for noncircular urban areas, complete sections of the distribution might be missing; for example, a city with a pie slice missing might be observed as one with a much lower maxima line over the range $\rho' > C$ than the one shown on the chart.

If in the process of measurement a false center is selected and accepted as true, the hypothesis (1) would be rejected even though true. Observations would be interpreted as contradicting the hypothesis (1), first because the greatest densities would not occur near $\rho' = 0$, i.e., near the false center, and, second, because scatter would occur (except very rarely). This seems to be what other investigators have tended to do. Colin Clark (1951) states, "Then (except in the central business zone)—$y = Ae^{-bx}$." He also wrote (1958): "Density falls as we proceed outwards from the centre of a city (subject only to the qualification that at the very heart of the city most of the space will be occupied by commercial buildings, and residential density will be low)." Duncan (1957) remarks: "...the central area is essentially a non-residential district whose resident population density usually is far below that of the contiguous areas." Muth (1960) measuring log Density on distance regressions excludes central business district census tracts for theoretical reasons which presumably are in part based upon fact. Reinhardt (1950) discards the hypothesis in favor of one which is sufficiently flexible to permit a relatively flat segment or a hook in the regression of log D on distance for the shorter distances from the assumed center.

It may be that these investigators are correct and the declining exponential hypothesis is false—at least for part of the distance range. Certainly, however, an hypothesis should not be rejected without critical testing in the face of observations which might be sufficiently consistent with it to confirm it.

With respect to scatter, investigators seem to be willing to accept it as consistent with the hypothesis (1). Clark (1958) uses an average density at a given distance, saying "fitting a line to the weighted data," i.e., "density of each individual ward or census tract as a function of its distance from the centre of the city," "proved to be impracticable." Muth (1960) with correlation coefficients squared ranging from 0.022 to 0.74 for samples of 25 census tracts in each of 46 cities, apparently does not conclude this contradicts the hypothesis (for part of the distance range) or requires reconsideration of the measuring procedure. Yet, the hypothesis (1) if true, indicates correlations should be high unless the center has been falsely located. And, if not high, then the first step in further study is to test the validity of the location of the center.

If in the process of measurement the center selected is recognized as a false one, then the true center can be located—under the assumptions we have made—either (a) by a trial and error process of changing the location or (b) by the solution of a set of equations obtained from 4 non-identical observations. As before, if a great deal is known, parameters can be determined with very few observations and without regard to the character of the sampling. Of course, many observations falling on the theoretical functional relation strengthen the conclusion and, of course, one which does not so fall invalidates the hypothesis.

Thus far, density, D, has been treated as a measurable characteristic of a person. I now wish to consider density, δ, as a measurable characteristic of an area. This apparently is the definition used by at least some other investigators. Corresponding to this change in definition of density, the definition of the coordinates, (ρ, θ), or (x, y) changes to that of location for an area.

Under these definitions, the hypothesis simply becomes (1) with δ in place of D and with (ρ, θ) locating a point within the area. It is accurate for infinitisimal areas, a close approximation for small finite areas, and, perhaps, only a crude approximation for larger finite areas.*

*Of course, it is theoretically possible to select points within areas which will satisfy a known functional relation. However, under observational conditions the location points are selected without knowing the functional relation.

Unlike the case for density as a variable attached to persons, however, the theoretical functional relation for density, distance, and direction specified by the hypothesis does not entail a probability distribution for the variables. Hence, in order to proceed at all, a probability distribution must be introduced into the theory or model.* As a minimum, it is necessary to have a conditional probability function.

Other investigators seem to have made the simple blanket assumption of random normally distributed log δ deviations. This is indicated by the least squares or regression line approach with log δ as the dependent variable and ρ' as the independent variable. Further indications are provided by the failure to discuss variation in the size of areas, variation in frequencies from area to area, and effects of non-circularity of a city.

I am not, at this time, prepared to say the assumption is wrong, but I do wish to probe a bit into the problem.

Measurement of densities is, of course, essentially a counting of frequencies if the areas are assumed known. Thus we have

$$\delta_i = f_i / A_i \tag{10}$$

Substituting in hypothesis (1), we have

$$\log f_i - \log A_i = \log \delta_0 - \frac{2}{\mu} \rho_i \tag{11}$$

as our functional relation. If we assume the area measurements are without error, then the probability distribution we seek is that of

$$P\left(\log \frac{f_i'}{f_i} \middle| \rho_i\right) \tag{12}$$

where, as previously noted, f_i' is the observed frequency and f_i, the theoretical frequency for the i^{th} area. Obviously, P depends upon how the f_i' are generated. But looking at (1) and recognizing that f_i' may be a binomial variable, one wonders whether the presumption of random sampling of persons is not to be used, at least as a beginning. In short, it appears that, even though we started by trying to avoid this presumption, we have come back to it.

The regression technique may provide satisfactory estimates of μ under the assumptions that f_i' is a random binomial variable, $f_i = f_j$ and any dependence between f_i' and f_j' can be neglected. Furthermore, if the f_i are not equal but each is small relative to $\sum f_i$, approximate adjustments may be possible by weighting by a simple function of A_i and $1/f_i$ (in practice $1/f_i'$).

However, the regression procedure may not provide satisfactory estimates of log δ_0. And, of course, estimation of the parameters of location is not included in the regression technique. In fact, it is precluded by the use of a conditional probability function.

The shortcomings of the regression method suggest the possibility of using the multinomial distribution--the standard for random sampling from a multi-celled universe. The mathematical expression, however, does not appear to be tractable --the term for the theoretical probability of an observation in the i^{th} cell being

$$\frac{1}{2\pi}\left(\frac{2}{\mu}\right)^2 e^{-\frac{2}{\mu}\rho_i} A_i k$$

where

$$\rho_i = \sqrt{(x-c_x)^2+(y-c_y)^2} \tag{13}$$

$$= \sqrt{\rho'^2 + c^2 - 2c\rho'\cos(\phi-\theta')}$$

and k is a factor of proportionality to obtain unity as the sum of the terms. Thus, the complete multinomial expression is obviously not easily applied to empirical data.

I have exhausted the time allotted to me. The results I have presented are, of course, incomplete and perhaps only of limited usefulness. I felt it was best to present them now, since there is no assurance I will have the opportunity to do further work. Perhaps others will and it is my hope that my work to date, such as it is, will be of use to them.

*Unless one wishes to accept least squares per se, or some similar procedure.

Selected References

Blumenfeld, Hans, "On the Concentric-Circle Theory of Urban Growth," Land Economics, Vol. 25, No. 2 (May 1949), pp 209-212.

Clark, Colin, "Urban Population Densities," Journal of the Royal Statistical Society, Series A, Vol. 114 (1951), pp. 490-496.

-----, "Urban Population Densities," Bulletin of the International Statistical Institute, Vol. 36, No. 4 (Actes de la 30e session de l'Institut International de Statistique, Stockholm, 1958), pp. 60-90.

Duncan, Otis Dudley, "Population Distribution and Community Structure," Cold Spring Harbor Symposia on Quantitative Biology, Vol. 22 (1957), pp. 357-371.

Muth, Richard, "Spatial Structure of the Housing Market," (In process, 1960).

Reinhardt, Forrest Rex, A Test of Hypotheses Specifically Related to the Cross-Sectional Distribution of Population Densities of Cities (In partial fulfillment of the Degree of Master of Arts in Sociology), University of Minnesota, September 1950, pp. 104 (typewritten).

Stewart, John Q., "Suggested Principles of 'Social Physics,'" Science, Vol. 106 (1947), pp. 179-180.

-----, and William Warntz, "Physics of Population Distribution," Journal of Regional Science, Vol. I, No. I (1958), pp. 99-123.

Winsborough, Halliman H., A Comparative Study of Urban Residential Densities, (In process, 1960).