Professional Geographer, 33(3), 1981, pp. 341–349

THE SIGNIFICANCE OF POTENTIAL-DENSITY REGRESSIONS

Michael F. Goodchild
The University of Western Ontario

Ralph F. Milliff and Scott M. Davis
University of California, Santa Barbara

Strong correlations have been observed between potential indices and the densities of spatial variables. The conventional null hypothesis of bivariate regression is inappropriate for testing their significance. A randomization test is proposed and is applied to 1975 US population data by state. The resulting relationship has a significant correlation, but its slope could occur frequently under the null hypothesis. The correlation is shown to be related to the spatial autocorrelation of densities by constructing arrangements with prescribed values of the modified Moran index.

THE purpose of this paper is to shed some light on the interpretation of empirical correlations between potential indices and densities. Suppose that an area is partitioned into a number of regions, and for each some spatially aggregated statistic P is known, for example, the population. The density of region i is calculated by dividing by the area A_i, that is, $D_i = P_i/A_i$. Although the example of population will be used in this paper, a large number of other measures are mentioned in the literature, particularly gross income (see for example [9]).

Potential is defined as a distance-weighted sum,

$$V_i = \sum_k P_k/d_{ik},$$

where d_{ik} is the distance from region i to region k. We shall not be concerned in this paper with the basis for this index, which has recently been reviewed by Sheppard [6]. But two practical comments should be made. First, if distance is measured between fixed points in each region, centroids for example, then the contribution to the sum is infinite when $k = i$. This self-potential problem has been dealt with in a number of ways. Perhaps the least arbitrary is to assume that the density within each area is constant and that the area is circular. Then the contribution for $k = i$ can be calculated by assuming the region to be composed of an infinitely large number of infinitely small pieces and integrating, to obtain $d_{ii} = \frac{1}{2}(A_i/\pi)^{\frac{1}{2}}$. Warntz [12] attributes this development to Court.

The second problem is that the calculated value of V_i clearly depends on the configuration of regions. It is presumed (see for example [11, p. 94]) that V_i is a good approximation to the integral

$$V(\boldsymbol{x}) = \int \frac{\rho(\boldsymbol{y})}{|\boldsymbol{x} - \boldsymbol{y}|} dA(\boldsymbol{y}),$$

where ρ is a continuous population density surface, and **x** and **y** are position vectors, and the integral is over the study area. But this has never been evaluated directly. Craig discussed some of the effects of region size on V_i and concluded that they could not be dismissed [2].

Despite these problems, a considerable amount of largely inductive research has been carried out using V_i estimates. The focus of this paper is on observed correlations between V_i and various spatial densities. Stewart was perhaps the first to observe a relationship of this type when he regressed log V_i against the log of the rural population density of the United States and obtained a regression slope very close to 2 [7]. He found this basic form to hold for a number of US Census years. Warntz examined the relationship between income potential and income density and found an approximately cubic law to hold consistently from 1880 to 1976, based both on states and counties (see [12] for a review). In the social-physical paradigm, with its emphasis on dimensional analysis and the fundamental quantities of social processes, the slopes of log-log regressions have great significance, particularly if they are close to integers.

The main concern in this paper is with doubly logarithmic regressions between pairs of indices, potential and density, which are both based on the same quantity P_i, for example, between population potential and population density or income potential and income density. When the quantities are different, as in regressions between rural population density and total population potential, the problem is more complex. Comments will be made on this second class toward the end of the paper.

Examples used in the paper will be based on the United States population statistics for 1975 by state, for the forty-eight coterminous states. Regression of log density against log potential gave a slope of 2.741, with a Pearson correlation of 0.899. This result is similar to those obtained by several workers using different time periods. The convention of regarding log potential as the independent and log density as the dependent variable will be followed to be compatible with the bulk of the literature.

Very little has been written on the interpretation of this kind of result. There is no theoretical basis for the analysis, except for a tenuous analogy to thermodynamics and gravitation. The standard null hypothesis of a bivariate regression is that the cases are sampled from a bivariate normal distribution with zero covariance. It is possible to test this using either the correlation coefficient or the regression slope (see, for example, [11, p. 97]). But this null hypothesis is clearly inappropriate in this case, as both variables have been computed from the same set of P_i statistics. We cannot, therefore, ignore the possibility that such relationships might occur by chance under a more appropriate null hypothesis.

The meaning of the relationship has received surprisingly little comment, given its observed strength, and it is remarkable that it appears to have led to so little further work. Warntz writes of the income relationship that

> Its theoretical underpinnings are not yet fully established, although evidence suggests that this form results from the interplay of success and failure in the myriad locational decisions made in competitive economic systems. This resulting pattern is then perhaps the give and take profit maximizing one for economic locations [12].

The potential index is a distance-weighted sum and can be visualized as an operator or filter applied to the density surface; because weight decreases with distance, V_i resembles local density values more than distant ones. The smoother the density surface, then, the higher we might expect the log-log correlation to be with potential [10, p. 16]. Thus one interpretation of the observed correlation is that it indicates a degree of smoothness in the density surface. Hirst ascribed the relatively low log-log correlation (about 0.73) between income potential and income density in Tanzania to a lack of smoothness [4, p. 280].

However, if V_i resembles local densities more than distant ones, it clearly most resembles the density of the region itself, D_i. This suggests that the relationship will occur to some extent whatever the spatial distribution of the underlying statistic P_i, and whatever is measured by the statistic. Before any meaning can be attached to the relationship, it is clearly necessary to evaluate this possibility. That is the main purpose of this paper.

The Problem

Doubly logarithmic regression between density and potential implies the model

$$\log P_i/A_i = a + b \log V_i,$$

which is equivalent to $P_i/A_i = a' V_i^b$, and where a, b and a' are constants. Writing $y = \log P_i/A_i$ and

$$x = \log V_i = \log[2P_i(A_i/\pi)^{-1/2} + \sum_{j \neq i} P_j/d_{ij}]$$

gives the conventional linear model $y = a + bx$. The estimated correlation and slope both depend on the covariance of x and y, $\sum (x - \bar{x})(y - \bar{y})/n$, which is a complex function of the populations, areas, and distances and does not simplify readily. But it is clear that, of the two terms in the expression for V_i, the first, which depends on P_i, will contribute to the covariance of x and y whatever the spatial distribution; the second, which depends on the populations of all other zones weighted by the inverse of their distances, will contribute most to the covariance when the density values are spatially autocorrelated, or smooth. The relative importance of these two effects is tested by randomization.

Randomization

Smoothness is a property of the spatial arrangement of the population densities. If the densities are rearranged, or shuffled, among the states, any property of the spatial arrangement will be destroyed but all other properties, such as the mean and standard deviation, will remain the same. Thus if the relationship between density and potential is a function of the smoothness of the density surface, it should be destroyed by a random permutation of the density values.

Let $R(i)$ denote the position of state i in a random permutation of states. Then $D_{R(i)}$, $i = 1 \ldots n$ is a random permutation of the original list D_i, $i = 1 \ldots n$. The new population of state i can be calculated as $D_{R(i)}A_i$, in other words, the population state i would have if it had the density of another, randomly chosen state.

In order to test whether the relationship is indeed a result of the spatial arrangement of densities, new potentials were calculated as

$$V_i' = \sum_k D_{R(k)}A_k/d_{ij}$$

and their logs regressed against log $D_{R(i)}$. One hundred independent randomizations were made to obtain an estimate of the distribution of both slope and correlation. In effect, these are the distributions of two test statistics under the null hypothesis that the observed density values are arranged randomly on the map. The random permutations were obtained by generating forty-eight random numbers and using the sequential position of the ith smallest number as the index $R(i)$.

The results are shown in Table 1 in the form of the means and standard deviations of the test statistics under the null hypothesis, and the number of simulation runs in which values more extreme than the real values were observed. We can use the latter to make a conservative estimate of the appropriate α level in the following way. Let p be the

TABLE 1
RANDOMIZATION RESULTS

	100 Simulations										Observed	
	Slope					Correlation						
	$\hat{\mu}_b$	$\hat{\sigma}_b$	n	α_1	α_2	$\hat{\mu}_r$	$\hat{\sigma}_r$	n	α_1	α_2	Slope	Correlation
Randomized Density	2.570	0.540	33	0.41	0.33	0.537	0.078	0	0.03	<0.01	2.741	0.899
Randomized Population	3.106	0.441	20	0.27	0.20	0.813	0.040	0	0.03	<0.01		

n = number of times simulated statistic was more extreme than the observed value.
α_1 = conservative estimate of α.
α_2 = maximum likelihood estimate of α.

probability that any one simulation run will give a value of a test statistic more extreme than the real value. Suppose that such values are observed for a total of n out of the N runs. Then the probability of obtaining n or fewer such values is given by the appropriate cumulative term of the Binomial distribution

$$\sum_{r=0}^{n} \binom{N}{r} p^r (1-p)^{N-r}.$$

p will be a conservative estimate of α when this probability is 0.05. When $n = 0$, we can write the expression for α directly as

$$\alpha = 1 - (0.05)^{1/N}.$$

This estimate of α will be conservative because it is derived by finding that value of p for which no more than 5 percent of experiments, each involving one hundred runs, would give the observed value of n. The maximum-likelihood estimate of α is simply n/N; this value is also shown in Table 1.

Table 1 likewise includes results for a randomization of population values rather than densities:

$$V_i' = \sum_k P_{R(i)}/d_{ik},$$

with the regression between log V_i' and log $P_{R(i)}/A_i$. As a null hypothesis, a random rearrangement of populations rather than densities does not seem to be as appropriate; these figures are therefore mainly for completeness.

The conclusions to be drawn from Table 1 are as follows. In the case of slope, it is clear that a value of 2.741 is quite reasonable under the null hypothesis. This means that no significance can be attached to its appearance in reality. A slope close to 3 is not an indication of social or economic spatial processes at work on the demographic landscape, as it can arise as easily from a random rearrangement of densities. In effect, we should fail to reject the null hypothesis.

However, the real correlation differs significantly from those obtained under density randomization. The conservative estimate of α is 0.03, but the maximum-likelihood estimate is much smaller. We can conclude that, although no meaning can be attached to the regression slope, the observed correlation is significant at the 0.03 level and reflects some aspect of the spatial arrangement of densities that is destroyed by randomization. Note,

however, that the simulated correlations are substantially different from zero, which is the expected value under the conventional null hypothesis for regression.

Spatial Autocorrelation

It appears from the previous section that correlation is controlled by some aspect of the spatial arrangement of density related to smoothness. In this section we discuss an experiment to explore in greater detail this relationship between smoothness and correlation.

A convenient way to measure the smoothness of a set of values for arbitrarily shaped regions is through the Moran autocorrelation statistic as modified by Cliff and Ord [1]:

$$I' = n \sum_{i,j} w_{ij}(x_i - \bar{x})(x_j - \bar{x}) / \sum_{i,j} w_{ij} \sum_i (x_i - \bar{x})^2,$$

where w_{ij} denotes the weight given to region j in relation to region i and x_i is the value in region i. The diagonal terms w_{ii} of the weights matrix are set to zero.

The choice of the weights matrix is fairly arbitrary, although weight must clearly be a decreasing function of distance. In this study w_{ij} was set equal to the reciprocal of d_{ij} to be consistent with the potential measure. On this basis we find that the autocorrelation of the US 1975 population density values by state was 0.409. The expected value of I' for randomly rearranged densities is approximately zero, whereas negative values correspond to a "checkerboard" or spatially antipersistent pattern in which high densities are juxtaposed with low ones.

Of the 48! possible arrangements of the state density figures, the vast majority have values of I' very close to zero. In order to study the relationship between I' and the regression of density with potential, an algorithm was employed to generate a particular permutation $R(i)$ with an autocorrelation as close as possible to a prescribed target [3]. The main elements of the algorithm are as follows:

(1) Make a random permutation.
(2) Select a random pair of states and compute the autocorrelation that would result if the values assigned to the pair were exchanged.
(3) If the new autocorrelation is closer to the target, make the swap.
(4) If the new autocorrelation is within a tolerance of the target, or if more than a limited number of unsuccessful swaps have been attempted, stop. Otherwise go to Step (2).

An advantage of the algorithm is that rearrangement does not affect any of the aggregate properties of the data, such as the mean and standard deviation. Thus all of these can be set in advance, independently of the search for a target autocorrelation. In this particular application the set of density values remained the same as the real set throughout the analysis, and only the permutation changed.

Clearly a large number of permutations would give autocorrelations within a tolerance of any reasonable target. The purpose of the initial random permutation and the random selection of a pair in Step (2) is to ensure that, as far as possible, the eventual permutation is randomly chosen from the set of possibilities and is not subject to any systematic effects.

The range of possible values of I' depends on the set of density values and the weights. Thus, although targets as low as -0.5 were specified, the minimum I' found by the algorithm was -0.143 and the maximum 0.603. The corresponding permutations are shown in Figure 1, using the interval-free method developed by Tobler [8]. It is clear that the minimum I' has been found by placing all values near to the mean in the west and placing all extremes in the east. The four highest values have been assigned to scattered eastern states and are surrounded by some of the lowest densities. The maximum I' has been generated by creating a smooth surface with all high values grouped in the northeast. One should note that the choice of this location cannot be attributed to any characteristic

Figure 1A

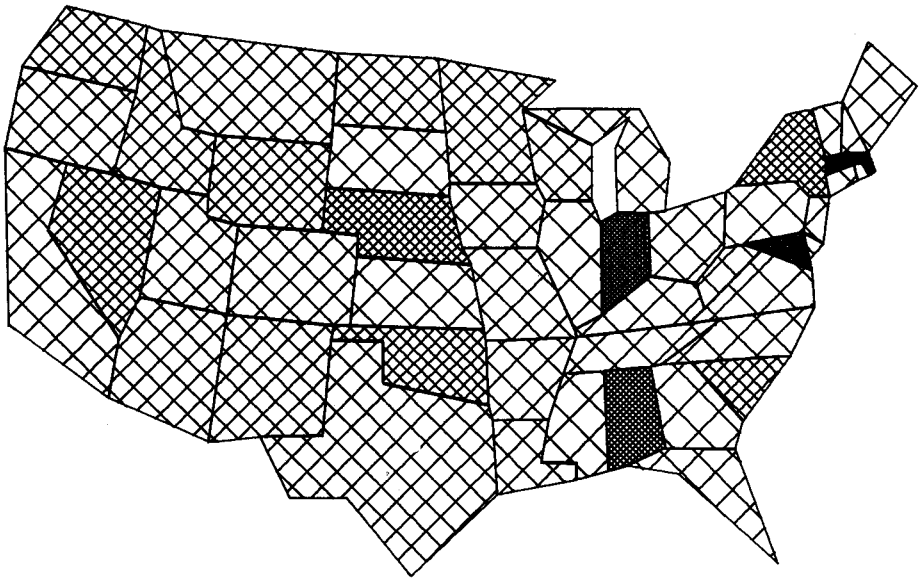


Figure 1B

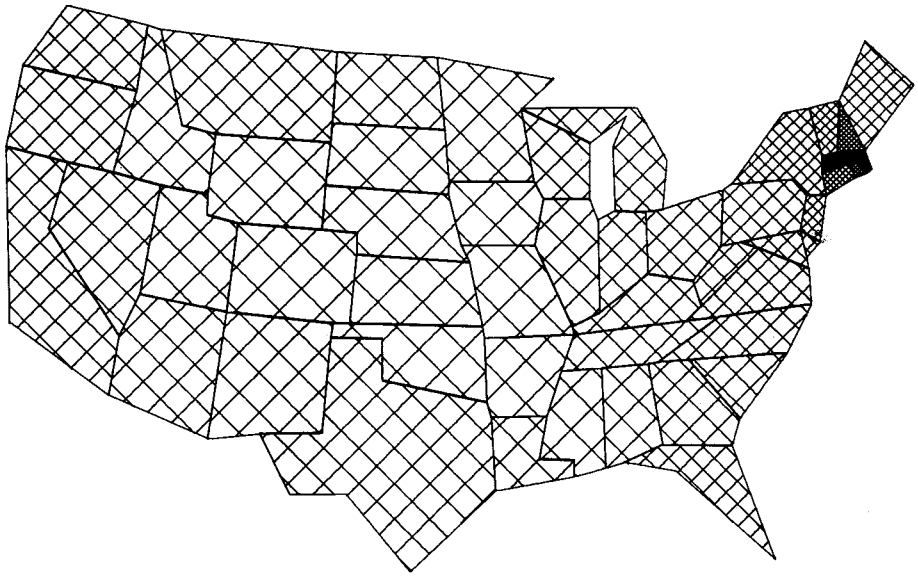
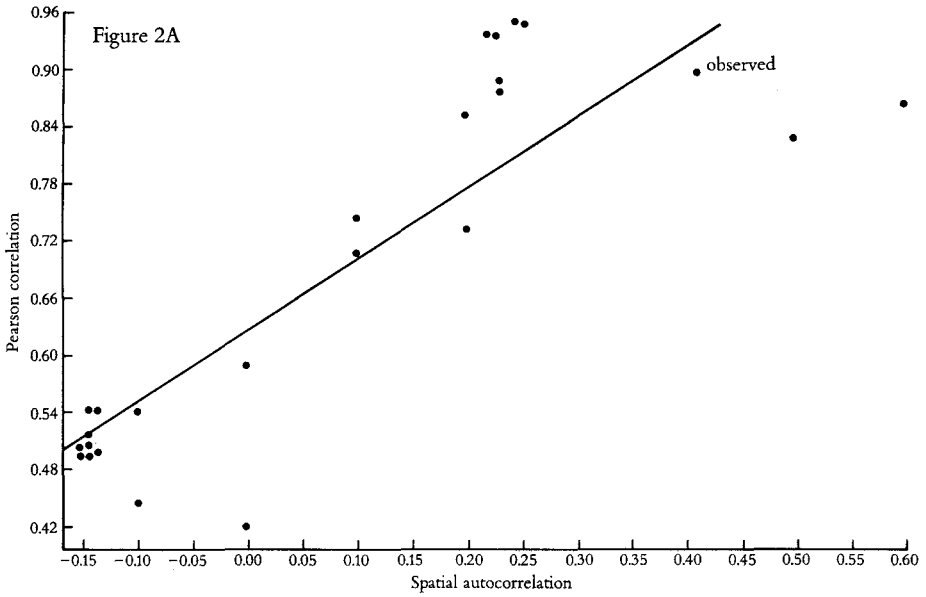


Figure 1. 1975 US state population densities permuted to extreme Moran spatial autocorrelations: (a) minimum = −0.143; (b) maximum = 0.603.



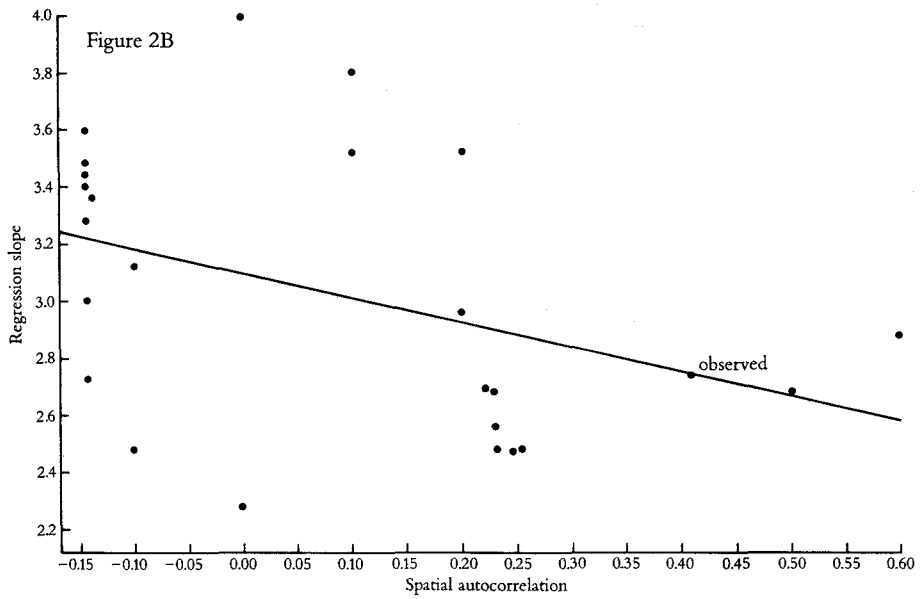


Figure 2. Scattergrams of test statistics for varying Moran spatial autocorrelations: (a) Pearson correlations, (b) regression slopes.

of the real pattern because this was totally destroyed in the initial shuffling in Step (1) of the algorithm. Instead, it results from the high weights that can be achieved over the short distances between centroids in this region. We note that, when several attempts are made to find the pattern with the maximum I', the resulting limit varies widely, depending on where the cluster of high values tends to develop. The algorithm yields the pattern of maximum I' only when the cluster develops in the northeast. On the other hand the minimum extreme is found with much greater consistency.

Twenty-four permutations were generated by making two replications each for targets of -0.5, -0.4 ... $+0.6$, to give a representative range of patterns. For each pattern, potentials were calculated and their logs regressed against the logs of densities, to give in each case a slope and Pearson correlation. These two statistics were then compared to the spatial autocorrelation (Figure 2). In the case of Pearson correlation the relationship was strongly positive ($r = 0.852$), indicating that the correlation between density and potential is a good indicator of underlying smoothness in the density surface, or vice versa. The observed correlation of 0.899 and autocorrelation of 0.409 for the U.S. 1975 data are consistent with the simulated results. On the other hand the relationship between regression slope and I' is weak ($r = 0.396$, not significant at the 0.05 level in a two-tailed test), confirming the previous conclusion that slope is not a significant index of spatial arrangement.

Conclusions

In regressions between potentials and corresponding densities, the conventional null hypothesis is inappropriate because both dependent and independent variables are linked to the same distribution of population. A nonzero correlation is to be expected for two reasons: first, because the greatest contribution to population potential is made by neighboring areas, which tend to have similar population densities; and, second, because the population of a place contributes to its own potential through the self-potential term. It is possible to test whether the regression is sensitive to particular aspects of the spatial arrangement of densities by randomization, using a null hypothesis that the observed densities are arranged randomly over the zones of the map. On this basis it appears that the correlation is significant but the regression slope is not. Correlation appears to measure a property of the spatial arrangement similar to that detected by the spatial autocorrelation coefficient. The slope, on the other hand, seems to be unaffected by spatial rearrangement of densities, which suggests that it is controlled by some aspatial aspect of the set of density values, such as the skewness or range. This is a possible topic for further research.

In general, the problem is an example of the inappropriateness of the conventional null hypothesis of bivariate regression when both dependent and independent variables share a common influence, in this case the population distribution, and is therefore a direct result of the definition of potential. Other examples can be found within geography and related disciplines; it would be even more inappropriate to use the conventional null hypothesis in a double-log regression between city size and city rank, because, in the case of the rank-size rule, one variable is derived directly from the other (see [5] for example). In this case it would be appropriate to replace double-log regression by a goodness-of-fit test of city size to the Pareto distribution.

Two points should be made about the generality of these results. First, they have been obtained specifically for 1975 United States populations by state, and they cannot be assumed to apply to all other situations, particularly at different levels of spatial resolution and when the self-potential problem is treated differently. Instead, significance should be evaluated directly by the randomization test. Second, several aspects of the comparison between regression parameters and spatial autocorrelation were arbitrary, including the definition of weights and the choice of the Moran index. It appears, however, that the

correlation between log potential and log density responds to the same general property of spatial arrangement as does the more conventional measure of spatial autocorrelation.

These conclusions have been based on regressions between densities and potentials derived from them. A second class of analyses, between potentials derived from one spatial variable, P_i, and some other density, Q_i/A_i, was mentioned earlier. Here the correlation between potential and density is combined with the correlation between the two densities. The significance of the spatial arrangement of P,Q pairs can be tested by repeating the analysis with randomized pairs. Randomization of P or Q alone will test whether the relationship between P and Q is significant.

Literature Cited

1. Cliff, A. D., and J. K. Ord. *Spatial Autocorrelation.* Pion, London, 1973.
2. Craig, J. "Population Potential and Population Density." *Area,* 4 (1972), 10–12.
3. Goodchild, M. F. "Algorithm 9: Simulation of Autocorrelation for Aggregate Data." *Environment and Planning A,* 12 (1980), 1073–81.
4. Hirst, M. A. "Telephone Transactions, Regional Inequality and Urban Growth in East Africa." *Tijdschrift voor Economische en Sociale Geografie,* 66 (1975), 277–93.
5. Malecki, E. H. "Growth and Change in the Analysis of Rank-size Distributions: Empirical Findings." *Environment and Planning A,* 12 (1980), 41–52.
6. Sheppard, E. S. "Interaction and Potential in Spatial Systems." *Ontario Geography,* 13 (1979), 47–60.
7. Stewart, J. Q. "Empirical Mathematical Rules Concerning the Distribution and Equilibrium of Population." *Geographical Review,* 37 (1947), 461–85.
8. Tobler, W. R. "Choropleth Maps Without Class Intervals?" *Geographical Analysis,* 5 (1973), 262–65.
9. Warntz, W. "Macrogeography and Social Science." *Geographical Review,* 48 (1958), 167–84.
10. ———. *Macrogeography and Income Fronts.* Regional Science Research Institute, Philadelphia, Pa. Monograph Series Number 3, 1965.
11. ———. "New Geography as General Spatial Systems Theory—Old Social Physics Writ Large?" In *Directions in Geography,* pp. 89–126. Edited by R. J. Chorley. Methuen, London, 1973.
12. ———. "Places of Birth, Education and Activity of American Leaders." *Ontario Geography,* 13 (1979), 3–24.

MICHAEL F. GOODCHILD is Professor of Geography at the University of Western Ontario. He obtained his Ph.D. from McMaster University and is interested in location-allocation methods, geographical data processing, and the application of statistics. RALPH MILLIFF is on the research staff of the Department of Geography, University of California, Santa Barbara, and is currently involved in research in oceanography. He recently completed a dissertation on a model of migration. SCOTT DAVIS is enrolled in the doctoral program in economics at Stanford University. When this paper was written he was a graduate student in the Department of Geography, University of California, Santa Barbara.