

Smooth Pycnophylactic Interpolation for Geographical Regions¹

Waldo R. Tobler²

ABSTRACT: Census enumerations are usually packaged in irregularly shaped geographical regions. Interior values can be interpolated for such regions, without specification of “control points,” by using an analogy to elliptical partial differential equations. A solution procedure is suggested, using finite difference methods with classical boundary conditions. In order to estimate densities, an additional nonnegativity condition is required. Smooth contour maps, which satisfy the volume preserving and nonnegativity constraints, illustrate the method using actual geographical data. It is suggested that the procedure may be used to convert observations from one bureaucratic partitioning of a geographical area to another.

KEY WORDS: Bivariate interpolation; Density estimation; Dirichlet integral; Numerical methods; Population density.

1. INTRODUCTION

The objective of this article is to clarify procedures for the preparation of a smooth map of a geographical distribution under the constraint that the original data arrive packaged in discrete collection regions. The latter situation is quite common in practice. One can, for example, obtain aggregate counts of individuals by state. From these data one might like to know how population density, a continuous quantity, varies over the particular portion of the earth. The assumption usually made is that the density within any individual reporting region is a constant, and it is implicitly asserted that this is optimal given that one lacks information to the contrary. For a single isolated region this assumption appears plausible, but for an interconnected set of regions it seems dubious. A common fact of geography is that places influence each other. This mutual influence of places can be interpreted mathematically, and one can exploit this geographical structure in order to enhance statements about places on the basis of a familiarity with events at nearby places.

The initial assumption is that there exists a density function, call it $Z(x, y)$, which is nonnegative and has a finite value for every location x, y in the domain of concern. As a matter of notation, I distinguish among the several regions by the use of a single subscript. Thus R_i means the i^{th} region, H_i denotes the value observed in this region, and A_i is the area of the region in square kilometers. The data regions are conveniently defined by polygons with a finite number of vertices using geographical coordinates. Thus the boundaries of the 3,077 counties of the contiguous United States can be described by 46,142 ordered latitude and longitude pairs, available on less than two meters of magnetic tape. The state boundaries are well described by 15,296 points. The value observed in each region is assumed to be a count or enumeration; H_i is therefore a nonnegative finite integer. The density function to be found must have the pycnophylactic (mass preserving) property defined by

$$\iint Z(x, y) dx dy = H_i$$

for all regions. The ellipsoidal shape of the earth is here ignored, and I assume an equal area map projection.

One function that exactly matches the requirements is the uniform density, $Z(x, y) = H_i / A_i$ if the location x, y is in H_i . Such a function is shown in the accompanying perspective diagram (Figure A), where the regions are states and the observations are numbers of people resident in

each state on a particular day. Diagrams of this type are usually referred to as bivariate histograms. It is apparent that a contour map of this function would not be smooth. Nor do adjacent regions influence each other in the construction. My objective is to obtain a smooth density function, representable by contours.

2. PREVIOUS WORK

There is cartographic literature on this topic, for example, Robinson and Sale (1969, pp. 141-170), in which the resulting diagrams are known as isopleth maps. The method of construction of such maps was outlined in 1845 by Lalanne as follows (Robinson 1971):

“Suppose, in effect, that one partitions the territory of a country into a large number of sufficiently small parts so that it would provide a division as extensive as the communes of France; that at the centre of each of these divisions one raises a vertical, proportional to the specific population, or in other words, to the number of inhabitants per square kilometre in the territory of the commune in question; that one joins the extremities of all these verticals with a continuously curved surface, and finally one projects on a map, at a convenient scale, the contours traced on that surface which correspond to equidistant integral elevations: One will thus have lines of equal specific population and one will be able to observe the series of points along which the population is 30, 40, 50, ..., 100 inhabitants per square kilometre.”

The first population density isopleth map known was made in this manner by Ravn and published in 1857 (Robinson 1971). Thus the method of construction currently in use has not changed in more than 100 years. The value H_i/A_i is assigned to the geographical center of each of the regions, and the isopleths are drawn as if these values were isolated spot heights taken from a topographic surface. The mass-preserving property is generally not mentioned, an exception being Schmid and MacCannell (1955), who assert that this method yields approximately correct volumes. This latter idea perhaps stems from the conjecture that if the density in each region is linear, $Z_i = a_i + b_i x + c_i y$, with x, y in R_i and $Z_i \geq 0$ in R_i then the pycnophylactic condition is satisfied under arbitrary rigid “tiltings” of the plane Z_i about the geographic center of gravity, when this location is assigned the fixed density H_i/A_i . An isopycnic map could be made from such a piecewise linear density function, but it would consist of straight contour lines with jumps at the boundaries between polygons. Probably the most common technique is to connect centroids by a triangulation and then to construct a tent like surface from the observations H_i/A_i . For details see Schmid and MacCannell (1955). The first derivatives of the resulting contour maps have discontinuities along the triangulation lines instead of at the polygon edges. One might minimize these kinks, but it seems more attractive to search for a continuous and everywhere differentiable density function. Brooks and Carruthers (1953, pp. 162-165) do consider mass preservation, but they treat the unnatural case in which the polygons are rectangular in shape and do not recognize that more than one mass preserving function can exist. Interpolation and contouring from values given at point locations is a much studied problem (for reviews, see Crain 1970; Schumaker 1976; Lawson 1978; Tobler 1979), but this literature is largely irrelevant to the present discussion.

Nordbeck and Rystedt (1970) cover the case in which individual people are directly observed at coordinate locations. A rectangular kernel - the “floating grid” of Schmid and MacCannell (1955) - is then used to obtain a continuous and differentiable density function (also see Degani

and Porter 1977). This is just an elementary version of techniques described in the statistical literature (e.g., Rosenblat 1956; Parzen 1962; Bartlett 1963; Cacaullos 1966) to obtain empirical probability densities and takes advantage of the fact that in Sweden data are often publicly available with the geographical coordinates of individual houses. When attempting to estimate geographical population densities on the basis of direct observations of individuals, one should base the kernel on empirical evidence describing the activity fields of people as given by Hägerstrand (1957, 1967) for example, rather than simply assume convenient mathematical forms for these kernels. One should also recognize that these geographical fields are neither spatially homogeneous nor isotropic. In the present instance we do not have observations on individuals, only spatial aggregates. Thus the task is closer in spirit to the visual information processing problem of enhancing a picture that has been blurred by aggregation within spatial regions, as discussed by Harmon and Julez (1973) for square polygons. Thus the problem considered is to attempt to produce smooth maps directly from the aggregate data.

3. AN APPROACH

The following visualization may be helpful. Imagine that Figure A consists of blocks of clay, each state being represented by a different color, and that the masses of clay are proportional to population. We now wish to sculpt this surface until it is perfectly smooth, but without allowing any clay to move from one state to another and without removing or adding any clay. This physical picture is a reasonable approximation to the mathematical method proposed. The real analytical difficulty seems to lie in describing realistic geographical polygons such as Florida, Michigan, and Cape Cod, all with prespecified mathematical basis functions. Geographical regions are frequently made up of several disjoint pieces, islands, or are multiply connected, containing lakes. "Cuts," sets of zero measure, are used in the polygon definitions. These practical considerations make it difficult to apply directly the elegant histospline technique of Boneva, Kendall, and Stefanov (1971) or the extension by Schoenberg (1973), both of which work so well for simple rectangular polygons. A solution for regular polygons is of little geographical importance. Because of these mathematical difficulties, an approximate numerical approach is proposed. One can use a system of finite elements (Prenter 1974; Mitchell and Wait 1977), or one can superimpose a fine mesh of equally spaced points over the domain and approximate a solution at these mesh points. I have adopted the latter approach. The fineness of the lattice must be sufficient to ensure that every polygonal region is represented by at least one, and preferably several, mesh points. Improved rules for the choice of the mesh size would be helpful.

After finding the density values at the superimposed lattice of points, using the method described in the following paragraphs, a density map can be drawn. The values at the lattice points are labeled z_{ij} , where the double subscripts i and j represent the row and column indices for the lattice, and the notation Z_k is used if the lattice point i, j is in region k . The pycnophylactic condition can then be enforced by requiring that the Riemann sum

$$\Delta x \Delta y \sum_k Z_k = H_k,$$

is preserved, where Δx and Δy represent the lattice spacing. This method of accumulating densities is appropriate if one displays the resulting values in discrete form on a line printer or otherwise as a grey scale image. But to display isopycnic lines as contours, the values z_{ij} should be regarded as a sampling of the function $Z(x, y)$. Constructing contours from a lattice is usually done by using linear interpolation (Cottafava and LeMoli 1969) so that the trapezoidal rule should be used to enforce the volume condition. This has a curious consequence. If all the

regions satisfy the pycnophylactic constraint and all lattice points also satisfy the nonnegativity constraint, $z_{ij} \geq 0$, then the lattice points immediately adjacent to a region of zero content must have zero density. Otherwise, because the contribution of each lattice point depends on how rapidly the surface slopes toward the neighbors, there is a small wedge of volume into the region of zero content. A somewhat similar effect, of opposite sign, was observed by Boneva et al. and becomes even more troublesome if Simpson's rule, or more refined methods (Davis and Rabinowitz 1967), are used for the quadrature. The effect can be lessened by the introduction of interregional boundary points between the nodes of the mesh. As a practical matter, the lattice is assumed fine and the effect is small; thus the crudest form of integration suffices.

A smooth function, intuitively, is one that has few oscillations, or on which neighboring points have similar values, or one that has a small rate of change in all directions. Adopting this last definition, in which the partial derivatives are small, it is natural to minimize the sum of the squares of these partial derivatives, that is, minimize $\iint [(Mz/Mx)^2 + (Mz/My)^2] dx dy$. This equation is known as Dirichlet's integral and has been studied extensively. Without the pycnophylactic and nonnegativity constraints, the minimum is given by Laplace's equation (Kantorovich and Krylov 1958, pp. 246 et seq.): $M^2z/Mx^2 + M^2z/My^2 = 0$. The lattice approximation to this last equation requires that the value at any lattice point approach the average of its neighbors. An even stronger condition requires that the averages of overlapping neighborhoods be similar to each other, or that some higher order of partial derivative at each point has the same value as the average of the neighboring partial derivatives of the identical order. If the derivatives are smooth, then the function must certainly be smooth. Thus one might be led to a minimization of the linearized version of the curvature of the surface $Z(x, y)$, that is, simplifying somewhat (cf. Weinstock 1974; Aleksandrov, Kolmogorov, and Lavrent'ev 1969), minimize

$$\iint [M^2z/Mx^2 + M^2z/My^2]^2 dx dy$$

Without the present constraints the solution to this problem yields the biharmonic equation:

$$M^4z/Mx^4 + 2 M^4z/Mx^2My^2 + Mz^4/My^4 = 0,$$

which is often treated as providing a minimization of energy in the linearized theory of elasticity (Birkhoff and Garabedian 1961; Briggs 1974). This does not exhaust the possible definitions of smoothness; Birkhoff and DeBoor (1965) give another. Perhaps more important one should observe that these minimizations are all in mean square and over the entire domain of interest. They do not require that the maximum departure from smoothness at any particular point be minimized. The present approach is similar. Reasoning by analogy, either the Laplacian or the biharmonic equation can be taken as the basic smoothness criterion, and then it requires only a slight modification in order to incorporate the pycnophylactic constraint (see Appendix).

5. THE COMPUTATIONAL STEPS

The continuous solution to Dirichlet's equation involves subtle mathematical difficulties (Folland 1976), but these are not of concern here since the finite difference versions, in which one replaces the derivatives by difference expressions such as

$$M^2z/Mx^2 = (Z_{i,j+1} - 2 Z_{i,j} + Z_{i,j-1}) / \Delta x^2,$$

are not subject to these difficulties (Epstein 1962, p. 200). For a square lattice the finite difference approximations to Laplace's equation and to the biharmonic equation are simple and well known (Forsythe and Wasow 1960; Wachspress 1966; Birkoff 1972; Ketter and Prawel 1972). The computer solution generally proceeds by iteration; for these elliptical partial differential equations extensive discussions of convergence and stability can be found in the

literature (Parter 1965; Walsh and Young 1953; Young 1954). As can be seen in the technical appendix, the pycnophylactic version of the Dirichlet problem has a similar linear form, and similar behavior can be anticipated. This conjecture is reinforced by my computational experience to date. The non-negativity constraint is more challenging, and I have only an ad hoc rule for this ease. This seems to be a common problem in density estimation procedures (cf. Tapia and Thompson 1978).

My FORTRAN program begins by assigning the mean density H_i/A_i to each lattice point in R_i and then modifies this by a small amount to bring it closer to the value required by the governing partial differential equation, given as a relation between neighboring lattice points. The pycnophylactic condition is enforced by incrementing or decrementing all the densities within individual regions after each computation, subject to the condition $z_{ij} \geq 0$. In the current computer implementation, three passes through the entire lattice are required. The first compares the lattice values against the chosen smoothness criterion and suggests the amount and direction of change to be applied at each point. The second pass modifies the suggested changes to enforce the pycnophylactic and nonnegativity constraints. Finally, adjustments are applied to the values at all lattice points. This ends one iteration, after which the mathematical sculpting is repeated. These iterations cease when all adjacent lattice points satisfy the smoothness criterion within some tolerance. Standard convergence-hastening techniques (Young 1962) should be investigated, although I have not done so. I have no doubt but that other improvements might also be made in the computational procedure.

The specific computer steps occur in two separate programs, as follows:

Step 0: Preprocessing: The N regions are described as polygons of a limited number of vertices. The map projection coordinates of these vertices and their sequential order are loaded into the program, and a lattice of equispaced points is then superimposed on this computer stored geographical map. Each lattice point is in turn tested against each polygon for inclusion until a match is found; a so-called "point-in-polygon" subroutine is used.

The result of this program is a set of lattice points, each labeled with the identification number (1 to N) of the region to which it belongs. Lattice points belonging to no region of interest are assigned the label $N + 1$. The boundaries between regions, and to the exogenous area, are thus described implicitly, as a change of label between adjacent lattice points. The boundary resolution is that of the lattice; standard techniques would allow this to be improved (Ketter and Prawel 1972, pp. 335-343).

The sequence that follows describes the second program that takes as input the lattice identifications, the populations by region, and an upper limit on the possible number of iterations. The phrase "for all lattice points" should be interpreted as meaning for all lattice points for which the label is N or less. Since each lattice point is identified by region, it is also possible to cumulate values for regions while processing lattice points.

Step 1: For all lattice points: Compute the adjustment for smoothness,

$$\delta_{ij}' = -z_{ij} + .25(z_{i,j+1} + z_{i,j-1} + z_{i+1,j} + z_{i-1,j})$$

in the Laplacian case, underrelax $\delta_{ij} = .25 \delta_{ij}'$ and store the cumulative adjustments for each region $s_k' = 3_k \delta_{ij}$. A similar expression for δ_{ij}' can be derived for the biharmonic equation and is incorporated in the program as an option. Values near the border are treated somewhat differently, as discussed in Section 6.

Step 2: For each region: Compute a decrementing factor so that the average adjustment is zero

$$s_k = -s_k'/A_k$$

Step 3: For all lattice points: Add the smoothing adjustment to the lattice value unless this

would make the density negative, that is, if $(z_{ij} + \delta_{ij} + s_k) \leq 0$, then $z_{ij} = \max(z_{ij} + \delta_{ij} + s_k, 0)$. Next cumulate the resulting population for all the regions $H_k' = \sum z_k$.

Step 4: For each region: Compare the cumulated population with the initially given population, and save the average difference $l_k = (H_k - H_k')/A_k$. This is necessary because of the nonnegativity constraint in step 3.

Step 5: For all lattice points: Add the average population difference unless this would result in a negative population density, if $(z_{ij} + l_k) \leq 0$ then $z_{ij} = \max(z_{ij} + l_k, 0)$ and assign any residual to the lattice points of that region that have not yet been examined, that is, if $(z_{ij} + l_k) < 0$, then increase l_k in such a manner that the residual will be evenly distributed over the remaining lattice points in region k.

Step 6: go to step 1 or stop. The stopping rules include exceeding an input number of iterations, or when all adjustments satisfy $(\delta_{ij}')^2 < \epsilon$ where $\epsilon = .001$.

This ends the computer algorithm, except for details of output. The treatment is nonstandard because of its inclusion of the pycnophylactic constraint in steps 2, 4, and 5 and because of the nonnegativity constraint in steps 3 and 5. Step 5 is not entirely satisfactory, but seems self-correcting in the course of many iterations. In particular, it will not work at the last lattice point in a subregion. The small error is corrected by step 4 on the next iteration. The program also has an option to delete the nonnegativity constraint in cases for which a negative interpolated value is geographically meaningful. An example would be net migrations, some of which are positive and some of which are negative.

6. BOUNDARY CONDITIONS

Since I am in effect solving an elliptical partial differential equation I must supply boundary conditions. Whatever value one assigns to the outside of the domain will affect the measure of smoothness near the edge, and this effect then propagates inward, as already recognized implicitly in some of the earlier literature (Schmid and MacCannell 1955). Two types of boundary specification are possible, and both are easily programmed for a digital computer, even for realistic geographical shapes. In the first instance, one can specify a numerical value for lattice points along the edge of the domain; this is known as the Dirichlet condition. All lattice points that fall outside the polygonal regions might, for example, be taken to be fixed at a density of zero when dealing with an area bounded by water. The other available type of boundary constraint requires the specification of the rate of change of the densities across the boundary, the so-called Neumann condition. Of course, one can mix these constraints depending on the information available for the exogenous geographical areas. A simple spatial rate of change condition applied at the boundary would assert that the gradient vanishes at the edge of the region, that is, $Mz/Mn = 0$, where n is the normal to the boundary of the domain. One would of course like the determination of the boundary condition to be a part of the mathematical specification, that is, what boundary condition yields the absolute minimum of the functional, subject to the constraints? This is the so-called "natural" boundary condition of classical mathematical physics (Kantorovich and Krylov 1958) and leads to $Mz/Mn = 0$. The interior densities cannot be determined in the approach adopted here until one specifies the boundary condition. The computer program allows a choice of either zero on the boundary or a zero gradient at the boundary.

7. EXAMPLES

We are now ready to demonstrate with examples. The first two are such that the density is

known to decline toward the edge of the domain. In all cases the pycnophylactic condition has been enforced by using Riemann sums. The first demonstration is a non geographic test and uses frequencies sampled from two overlapping bivariate normal distributions. The particular data were also used in the discussion following the paper by Boneva, Kendall, and Stefanov (1971) and thus provide a direct comparison to that work. The 98 observations are first aggregated into 25 rectangular regions and then quantized to a 46 X 46 mesh, surrounded by an exogenous region one cell wide. Both the aggregation and lattice were chosen arbitrarily. Laplace's equation was then approximated by using 200 iterations for this 48 X 48 mesh, at an approximate cost of \$1 per 100 iterations. The contouring of the lattice uses only linear interpolation (Cottafava and LeMoli 1969). The two results, Figure C, demonstrate quite dramatically the difference between the alternate boundary conditions. The main shortcoming of my method in this example appears to be that the absence of observations in some cells is taken quite literally; the algorithm does not recognize the sampled nature of the data. Nevertheless, the two peaks are resolved, and the general agreement with the "correct" solution (cf. Boneva, Kendall, and Stefanov 1971, Fig. 3, p. 47) is tolerable.

A second and more realistic example uses the 1970 population figures for the 18 census tracts covering Ann Arbor, a city of approximately 100,000 people. The conventional choropleth map and bivariate histogram for these data are shown in Figure D. The tracts are next approximated by a mesh (schematized in Figure E) arbitrarily chosen to be 68 X 71 in size. Two hundred iterations using the biharmonic equation as the target and contouring by linear interpolation result in the density maps shown in Figures F and G. The effect of the alternate boundary conditions is not large in the interior of the region.

The data for the third and final demonstration are the 1970 populations *by state* for the contiguous United States. The densities have here been computed at the nodes of a 62 X 97 mesh, this size being sufficient to assign four lattice points to the smallest state. Thus Figure A is converted into Figure H, where the resulting values are shown as maps of level curves. Two versions are presented, using alternate boundary conditions, of an approximation to the biharmonic equation. Since much of the United States is bordered by water, a Dirichlet condition of zero density was first used adjacent to the boundary. This procedure creates two peaks in California (*sic*) and sets most of Nevada to a zero density. It has also combined Miami and Atlanta, moved Chicago southward, and created a barrel-shaped density for Michigan. Use of the Neumann condition $Mz/Mn = 0$ allows the cities to move closer to the edge, where the density drops sharply to zero outside the United States, and seems to yield a better fit, at least to my a priori expectations.

8. DISCUSSION

The example using the population of the United States is well suited to demonstrate some of the difficulties accompanying my approach. If one contemplates possible applications of the interpolated densities, it is imperative to ask whether these bear any resemblance to actual densities. In the present instance we also have available population by county and by finer geographical subdivisions. Suppose that the population density at a lattice point, assuming uniform densities within counties, is c_{ij} . Then we can make two comparisons, namely,

$$\sum_i \sum_j (c_{ij} - z_{ij})^2 \text{ and } \sum_i \sum_j (c_{ij} - d_{ij})^2 ,$$

where $d_{ij} = H_k / A_k$ denotes the density at a lattice point assuming uniform density within states, and z_{ij} is the smooth interpolated density. These comparison computations have not been performed, but a population "density" by county map is available (U.S. Department of the

Interior 1970, p. 241). It is fairly obvious that the approach described is an improvement over the constant density assumption, and the method of squared deviations should in principle allow one to judge whether the use of information from adjacent polygons is beneficial. But there seems to be an infinite regress here, since the smooth interpolation can always be applied to the finest subdivision possible. It would always be assumed that one has used the most detailed data available. Thus it is common practice to supplement census enumerations by using aerial photographs. In effect this provides a redefinition of the polygonal areas and does not constitute a real change in the problem. Eventually one reaches the level of the individual objects, and the definition of density itself becomes fuzzy. The problem is quite similar to the deblurring of photographs, in that one is attempting to invert a local accumulation process (Rosenfeld and Kak 1976, pp. 203-252). The amount of a priori information that one brings to such a situation decisively influences the quality of the result. And there seems to be no end to the possible additional detail that one might attempt to build into the algorithm. In the present instance the location of Chicago, say, might be specified by giving coordinates at which the density is to be a downward convex stationary point $Mz/Mx = Mz/My = 0$, $M^2z/Mx^2 < 0$, $M^2z/My^2 < 0$. Such additional information could easily be incorporated in a computer program. Another modification could be to allow the effect of transportation routes within the separate polygonal regions, allowing variable permeabilities in different directions. In the finite difference equations this would imply differential weighting of the neighbors, resulting in nonhomogeneous and anisotropic smoothing. These types of modifications are really too complicated to consider here, especially since it is not obvious that one would ever have the necessary empirical information. Perhaps one can discover a differential equation that describes geographical clusters of people, as suggested by Christaller's (1966) central place theory, and use this as the target, rather than borrow equations from mathematical physics. Such an equation, being based on geographical theory, should capture more of the phenomena.

A more fruitful set of variations, which I have not pursued, would seem to be along the following lines. One can assume that the original data contain some error. Then a modest amount of displacement of people from one region to adjacent regions might be allowed and an entropy function minimized (Frieden 1975; Pizer and Vetter 1968). Another variant would be to assume, or to estimate empirically, a spatial covariance structure for the interpolation (Kaula 1967) and then to incorporate this geographical autocorrelation in a procedure related to the method of Matheron (1971) or Moritz (1970). Alternately, the smoothing differential equation could be used as a target for a Monte Carlo simulation, assigning individuals to particular lattice points in a constrainedly random manner. Thus the different smoothness criteria could be interpreted as alternate ways of assigning occupancy probabilities to the lattice points (see Appendix). These several alternatives more closely resemble classical statistical density estimation in that a distribution of estimates would be obtained at each lattice point, rather than a single deterministic value. Such an attack would be of assistance in sampling situations, as already noted in the first example (Figure B and C).

9. CONCLUSION

The method described in this article allows the interpolation of values at a spatial mesh of arbitrary fineness from data given by irregular geographical polygons without any requirement for internal "control points" or "tent" functions. The isopycnic maps drawn from the mesh values are constructed to have the volume preserving property. A bivariate histogram can therefore be reconstructed exactly from the contour map simply by computing the "volume" under the

contoured surface within the irregularly shaped polygon. This is sufficiently important to be repeated: We can go from the contour map back to the original data! But the contours are not unique, and there is no way (short of a finer resolution in the original assembly of the data) to demonstrate the validity of the density at any particular point. The integrals over the original spatial packages are satisfied, and this result is as correct as possible for the conditions of the problem. Thus the mapping of the geographical arrangement of phenomena is improved. The critical assumption is that events in one geographical area influence those in adjacent areas. Concomitantly the great importance of events in exogenous regions translates into a necessity for the specification of boundary conditions.

The smooth volume-preserving density functions would also seem to offer an approach to the practical and frequently occurring problem of interconverting, or rendering compatible, data collected by different governmental agencies using completely distinct sets of geographical boundaries for the same part of the world (Markoff and Shapiro 1973; Crackel 1975; Ford 1976). One merely needs to reassign the lattice points, with their associated densities, to the alternate set of polygons. A simple addition over the lattice points contained in each new polygon then yields the approximate cumulant of the arrangement for that polygon. This result should provide a better conversion than one based on the less realistic assumption of constant densities within statistical data collection regions.

APPENDIX

1. An outline is given here to demonstrate the existence and uniqueness of the discrete Dirichlet problem with a pycnophylactic constraint. The approach is that suggested by Courant, Friedrichs, and Lewy (1928), to which the reader is referred for details and for a treatment of the boundary problem. For the sake of brevity the development here is given for only small examples, but there is no intrinsic difficulty in extension to larger problems.

Assume first that the result of the analysis is to be a one-dimensional histogram, with n cells of equal width Δ and heights Z . Then a measure of the smoothness of this histogram is the difference in heights between adjacent histogram bars, that is, $Z_{j+1} - Z_j$. For the entire histogram a natural measure of smoothness is the cumulative square of these individual values

$$T_0 = 3 \sum_{n=1}^{n-1} (Z_{j+1} - Z_j)^2.$$

which is a discrete approximation to the one-dimensional Dirichlet integral $\int (MZ/Mx)^2 dx$ as is easily seen if one sets $\Delta x = \Delta = 1$ and uses the forward difference approximation $MZ/Mx_j = (Z_{j+1} - Z_j)$. The pycnophylactic constraint requires that $\sum Z_k = H_i$, $k \in R_i$, for each region i . This is now added to the value to be minimized, using Lagrangian multipliers,

$$T = T_0 + 3 \sum_{i=1}^r \lambda_i (H_i - 3 \sum_{k \in R_i} Z_k).$$

Setting the partial derivatives of T with respect to each Z and each λ equal to zero yields a system of $n + r$ linear equations, where n is the number of lattice points and r is the number of regions. The system is of rank $n + r$ and thus has a unique solution.

When n is nine and r is two, for example, the system is small enough to be written out explicitly. Thus, putting the first four lattice points in region one and the remaining five in region two, the system becomes

| | | | | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|---|-----|---|----------------|----------------|
| 1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1/2 | 0 | Z ₁ | 0 |
| -1 | 2 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1/2 | 0 | Z ₂ | 0 |
| 0 | -1 | 2 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 1/2 | 0 | Z ₃ | 0 |
| 0 | 0 | -1 | 2 | -1 | 0 | 0 | 0 | 0 | 0 | 1/2 | 0 | Z ₄ | 0 |
| 0 | 0 | 0 | -1 | 2 | -1 | 0 | 0 | 0 | 0 | 1/2 | 0 | Z ₅ | 0 |
| 0 | 0 | 0 | 0 | -1 | 2 | -1 | 0 | 0 | 0 | 1/2 | * | Z ₆ | = 0 |
| 0 | 0 | 0 | 0 | 0 | -1 | 2 | -1 | 0 | 0 | 1/2 | | Z ₇ | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | -1 | 2 | -1 | 0 | 1/2 | | Z ₈ | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 1 | 0 | 1/2 | | Z ₉ | 0 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | λ ₁ | H ₁ |
| 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | | λ ₂ | H ₂ |

or $\mathbf{CZ} = \mathbf{H}$. The unique solution is $\mathbf{Z} = \mathbf{C}^{-1}\mathbf{H}$. In the present instance, if $H_1 = 8$ and $H_2 = 5$ then, $\mathbf{Z}^t = (2.20, 2.12, 1.96, 1.72, 1.39, 1.13, .93, .81, .74)$, $T = .3254$. If $H_1 = 5$ and $H_2 = 8$, then $\mathbf{Z}^t = (1.18, 1.21, 1.26, 1.35, 1.46, 1.56, 1.62, 1.67, 1.69)$, $T = .0401$, where the t denotes the transpose. It is clear that \mathbf{C}^{-1} is acting to distribute the population over the lattice points, and that \mathbf{C}^{-1} depends on the geography of the problem but not on the specific values in the vector \mathbf{H} . Thus this inverse need be calculated only once. But for the United States example given in the text it would be of larger size, involving 3,306 equations, which illustrates in part why iterative techniques are used to solve such sparse matrix systems. It also illustrates why I have used such a small example here.

The portion of \mathbf{C} near the diagonal has the form

$$\dots -1 \quad 2 \quad -1 \dots$$

This can be recognized as the coefficient form for the finite difference approximation to the one-dimensional Laplacian, $M^2Z/Mx^2 \approx (Z_{j+1} - 2Z_j + Z_{j-1})/\Delta^2$ aside from an unimportant change of sign. Thus it would have been possible to start directly from the Laplacian equation.

A simple two-dimensional example can be obtained by arranging the lattice points in a 3 by 3 array as follows:

$$\begin{matrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 2 & 2 & 2 \end{matrix}$$

where the numbers refer to the regional assignment to the two regions. Subtracting and squaring neighboring cell heights yields

$$T_0 = 3 \sum_{i=1}^{I-1} \sum_{j=1}^{J-1} (Z_{i,j+1} - Z_{i,j})^2 + 3 \sum_{i=1}^{I-1} \sum_{j=1}^{J-1} (Z_{i+1,j} - Z_{i,j})^2,$$

using row and column indices to distinguish between lattice points. I is the number of rows in the array, and J is the number of columns. Adding the pycnophylactic constraint gives

$$T - T_0 + \sum_{i=1}^r \lambda_i (H_i - 3 \sum_{k=0}^R R_k)$$

Thus the coefficient matrix \mathbf{C} becomes:

| | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|-----|-----|
| -2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1/2 | 0 |
| 1 | -3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1/2 | 0 |
| 0 | 1 | -2 | 0 | 0 | 1 | 0 | 0 | 0 | 1/2 | 0 |
| 1 | 0 | 0 | -3 | 1 | 0 | 1 | 0 | 0 | 1/2 | 0 |
| 0 | 1 | 0 | 1 | -4 | 1 | 0 | 1 | 0 | 0 | 1/2 |
| 0 | 0 | 1 | 0 | 1 | -3 | 0 | 1 | 0 | 0 | 1/2 |
| 0 | 0 | 0 | 1 | 0 | 0 | -2 | 1 | 0 | 0 | 1/2 |
| 0 | 0 | 0 | 0 | 1 | 0 | 1 | -3 | 1 | 0 | 1/2 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | -2 | 0 | 1/2 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |

Taking $H_1 = 8$ and $H_2 = 5$ yields the solution

| | | |
|-------|-------|-------|
| 2.240 | 2.060 | 1.956 |
| 1.740 | 1.306 | 1.170 |
| .996 | .816 | .711 |

and $T = 2.734$. The module surrounding -4 near the center of C is recognized as the finite difference approximation to the two-dimensional Laplacian. Alternate arrangements of the nine lattice points would result in slightly different versions of C . By measuring the “height” of lattice points along the edge of the 3 by 3 region one can also include the boundary in the minimization problem.

If some elements of H are negative, some histogram heights must also be negative. As indicated in the text, this may be meaningful geographically, but not for densities. Further, if, in the preceding examples $H_1 = 80$ and $H_2 = 5$, then Z contains negative numbers; there is as yet no nonnegativity constraint in the foregoing matrix equation. Such a constraint has been added to the computer program, as described in the body of the text. It should also be possible to solve this as a quadratic programming problem (Danzig 1963, pp. 490-497): Minimize T_0 , subject to

$$3 Z_k = H_i \text{ and } Z_k \geq 0.$$

2. In the Monte Carlo assignment of densities to lattice points, I envision a process as follows. Compute pycnophylactic densities as before; multiply these by the lattice spacing to obtain persons rather than densities, and then divide the number at each lattice point by the total population of the region containing that lattice point. The sum of all the numbers within each separate region is then equal to one, and these numbers can then be considered as probabilities. Linearly order the lattice points within each region, and then cumulate the probabilities within this ordering. Now sample a uniformly distributed random number as many times as there are people within each region, and each time assign an individual to the lattice point whose value in the cumulative distribution contains the random number in its span. The resulting geographical arrangement of individuals will be stochastic rather than smooth even though it was generated from a smooth arrangement of probabilities. Thus the resulting contours will have a more realistic appearance while still satisfying the nonnegativity and pycnophylactic constraints. This also suggests a method of producing smooth dot maps.

REFERENCES

Aleksandrov, A., Kolmogorov, A., and Lavrent'ev, M. (1969), *Mathematics: Its Content, Methods, and Meaning*, Vol. II (2nd ed.), Cambridge: MIT Press, Ch. VII, p 88.

- Bartlett, M.S. (1963), "Statistical Estimation of Density Functions," *Sankhya*, Ser A., 245—254.
- Birkhoff, G. (1972), *The Numerical Solution of Elliptic Equations*, Philadelphia: SIAM.
- Birkhoff, G. and DeBoor, C. (1965), "Piecewise Polynomial Interpolation and Approximation," in *Approximation of Functions*, ed. H. Garabedian, Amsterdam: Elsevier, 164—190.
- Birkhoff, and Garabedian, H. (1960), "Smooth Surface Interpolation," *Journal of Mathematical Physics*, 39, 258—268.
- Boneva, L., Kendall, D., and Stefanov, J. (1971), "Spline Transformations," *Journal of the Royal Statistical Society*, Ser. B., 33, 1—70.
- Briggs, J.C. (1974), "Machine Contouring Using Minimum Curvature," *Geophysics*, 39, 1, 39—48.
- Brooks, J.C., and Carruthers, N. (1953), *Handbook of Statistical Methods in Meteorology*, London: Stationery Office, 162—165.
- Cacaullos, T. (1966), "Estimation of a Multivariate Density," *Annals of the Institute of Statistics and Mathematics*, 19, 179.
- Christaller, W. (1966), *Central Places in Southern Germany*, trans. C. Baskin, Englewood Cliffs, N.J.: Prentice-Hall.
- Cottafava, G., and LeMoli, G. (1969), "Automatic Contour Map," *Communications of the ACM*, 12, 7, 386—391.
- Courant, R., Friedrichs, K., Lewy, H. (1928), "Über die Partiellen Differenzgleichungen der Mathematischen Physik," *Mathematische Annalen*, 100, 32-74.
- Crackel, J. (1975), "The Linkage of Data Describing Overlapping Geographical Units—A Second Iteration," *Historical Methods Newsletter*, 8, 3, 146—150.
- Cram, I. (1970), "Computer Interpolation and Contouring of Two-Dimensional Data: A Review," *Geoexploration* 8, 71—86.
- Danzig, G.D. (1963), *Linear Programming and Extensions*, Princeton, N.J. University Press.
- Davis, J., and Rabinowitz, P. (1967), *Numerical Integration*, Waltham, Mass.: Blaisdell.
- Degani, A., and Porter, P. (1977), "On Isopleths and Continuous Scanning-Isodensitron Entry," *Cartographic Journal*, 14, 1, 30—43.
- Epstein, B. (1962), *Partial Differential Equations*, New York: McGraw-Hill Book Co.
- Folland, G.B. (1976), *Introduction to Partial Differential Equations*, Princeton, N.J.: Princeton University Press.
- Ford, L. (1976), "Contour Reaggregation: Another Way to Integrate Data," *Papers*, Thirteenth Annual URISA Conference, 11, 528—575.
- Forsythe, G., and Wasow, W. (1960), *Finite Difference Methods for Partial Differential Equations*, New York: John Wiley & Sons.
- Frieden, B.R. (1975), "Image Enhancement and Restoration," in *Picture Processing and Digital Filtering*, ed. T. S. Huang, New York: Springer Verlag, 179—246.
- Hägerstrand, T. (1957), "Migration and Area," in *Migration in Sweden: A Symposium*, ed. D. Hannerberg, Lund Studies in Geography No. 13, University of Lund.
- Hägerstrand, T. (1967), *Innovation Diffusion As a Spatial Process* (Pred translation), Chicago: University of Chicago Press.
- Harmon, L., and Julez, B. (1973), "Masking in Visual Recognition," *Science*, 180—4091, 1194—1197.
- Kantorovich, L.V., and Krylov, V.J. (1958), *Approximate Methods of Higher Analysis*, Groningen: Noordhoff Interscience.
- Kaula, W. (1967), "Theory of Statistical Analysis of Data Distributed Over a Sphere," *Reviews*

of Geophysics, V, 1, 83—107.

Ketter, R., and Prawel, S. (1972), *Modern Methods of Engineering Computation*, New York: McGraw-Hill Book Co.

Lawson, C. (1978), "Software for C' Surface Interpolation," *Mathematical Software III*, ed. J. Rice, New York: Academic Press.

Markoff, J., and Shapiro, G. (1973), "The Linkage of Data Describing Overlapping Geographical Units," *Historical Methods Newsletter*, 7, 1, 34—46.

Matheron, G. (1971), *The Theory of Regionalized Variables and its Applications*, Fontainebleau: Cahiers du Centre de Morphologie Mathematique de Fontainebleau, No. 5, Ecole Superieure des Mines.

Mitchell, A., and Wait, R. (1977), *The Finite Element Method in Partial Differential Equations*, New York: John Wiley & Sons.

Mortiz, H. (1970), Eine Allgemeine Theorie der Verarbeitung von Schwermessungen nach Kleinsten Quadraten, Heft Nr. 67A, Munich: Deutsche Geodatische Kommission.

Nordbeck, S., and Rystedt, B. (1970), "Isarithmic Maps and the Continuity of Reference Interval Functions," *Geografiska Annaler*, Ser. B., 2, 92—123.

Parter, S. (1965), "On Estimating the 'Rates of Convergence' of Iterative Methods for Elliptic Difference Equations," *Transactions, American Mathematical Society*, 114, 320-354.

Parzen, E. (1962), "Estimation of a Probability Density and Mode," *Annals of Mathematical Statistics*, 33, 1065—1076.

Pizer, S.M., and Vetter, H. G. (1968), "Perception and Processing of Medical Radio-Isotope Scans," in *Pictorial Pattern Recognition*, eds. G.C. Cheng et al., Washington D.C.: Thompson Book Co., 147—156.

Prenter, P.M. (1974), *Splines and Variational Methods*, New York: John Wiley & Sons.

Robinson, A., and Sale, R. (1969), *Elements of Cartography*, New York: John Wiley & Sons.

Robinson, A. (1971), "The Genealogy of the Isopleth," *Cartographic Journal*, 49—53.

Rosenblatt, M. (1956), "Remarks on Some Nonparametric Estimates of a Density Function," *Annals of Mathematical Statistics*, 27, 832—837.

Rosenfeld, A., and Kak, A. (1976), *Digital Picture Processing*, New York: Academic Press.

Schmid, C.F., and MacCannell, E.H. (1955), "Basic Problems, Techniques, and Theory of Isopleth Mapping," *Journal of the American Statistical Association*, 50, 220—239.

Schoenberg, I. (1973), *Cardinal Spline Interpolation*, Philadelphia: SIAM, 115—119.

Schumaker, L. (1976), "Fitting Surfaces to Scattered Data," in *Approximation Theory II*, ed. G. Lorentz, New York: Academic Press, 203—268.

Tapia, R., and Thompson, J. (1978), *Non-Parametric Probability Density Estimation*, Baltimore: Johns Hopkins University Press.

Tobler, W. (1979), "Lattice Tuning," *Geographical Analysis*, XI, 1, 36—44.

U.S. Department of the Interior (1970), *National Atlas of the United States*, ed. A. Gerlach, Washington D.C.

Walsh, J., and Young, D. (1953), "On the Accuracy of the Numerical Solution of the Dirichlet Problem by Finite Differences," *Journal of the National Bureau of Standards*, 51, 343—363.

Wachspress, E. (1966), *Iterative Solution of Elliptic Systems*, Englewood Cliffs, N.J.: Prentice-Hall.

Weinstock, R. (1974), *Calculus of Variations*, New York: Dover Publications, Ch. 10, 228—260.

Young, D. (1954), "Iterative Methods for Solving Partial Difference Equations of Elliptic Type," *Transactions, American Mathematical Society*, 76, 92—111.

Young, D. (1962), "The Numerical Solution of Elliptic and Parabolic Partial Differential Equations," in *Survey of Numerical Methods*, ed. J. Todd, New York: McGraw-Hill Book Co., 380—438.

¹ © Journal of the American Statistical Association
September 1979, Volume 74, Number 367, 519-536
Application Section

² Waldo R. Tobler is Professor of Geography, University of California, Santa Barbara, CA 93106. This work was stimulated by the paper of Boneva, Kendall, and Stefanov (1971). On March 20, 1976, an alternate approach using the theory of cartograms was presented to the Workshop on Automated Cartography and Graphics in Epidemiology and Health Statistics, convened by the National Center for Health Statistics of the Department of Health, Education and Welfare.

The diagrams and analyses were done by using computer programs prepared by the author while at the University of Michigan. I am indebted to my colleague R. Leipnik for bringing the paper by Courant, Friedrichs, and Lewy (1928) to my attention. A listing of the computer program described may be obtained from the author.

COMMENT

Nira Dyn, Grace Wahba, Wing-Hung Wong
...(Several pages of comment omitted; relevant but lengthy)

REJOINDER

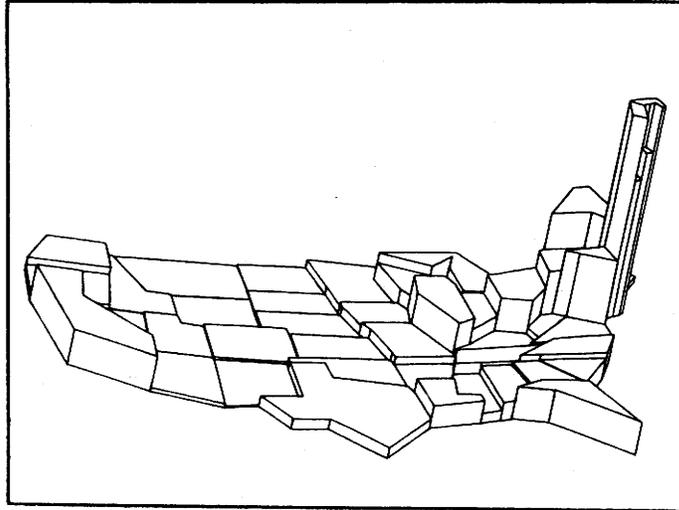
...(Several sections omitted)

The comments by Dyn, Wahba, and Wong specifically address only the formal minimization problem. This is important but it seems to me that other issues exist that might be addressed. For example, is "smoothness," however defined, a reasonable condition to impose on geographical data? The question of the intrinsic meaning of density has also bothered me. What meaning can numerical convergence to a mathematical limit have, for example, when there is no unambiguous definition of density for the raw, nonaggregated data? Thus I would like to propose the following. Assume that one has "points" located interior to a bounded portion of a plane. About each point define a polygon that has the property that all locations in the polygon are closer to the point than to any of the other points (i.e., Thiessen polygons, cf. Boots and Getis 1978, pp. 126-128). The generalization to higher dimensions seems to me to be immediate but not of geographical interest. Each such polygon will have a finite area. The reciprocal of this area gives the density associated locally with each point, and one can construct a bivariate histogram made up of these polygons and their "heights". The pycnophylactic sculpting is now applied to this histogram to obtain the smooth distribution of densities. The result of applying the pycnophylactic sculpting to any other aggregate data should converge to the density obtained in this manner from the raw data, as the aggregation regions decrease in size. This definition of density avoids the kernel size, shape, weighting, and orientation problems of the more usual density estimation procedures and is applied to an enumeration rather than to a sample of data. Of course it substitutes for these a certain ambiguity introduced by alternate definitions of smoothness and of boundary conditions. But these would be constants in any one problem.

REFERENCE

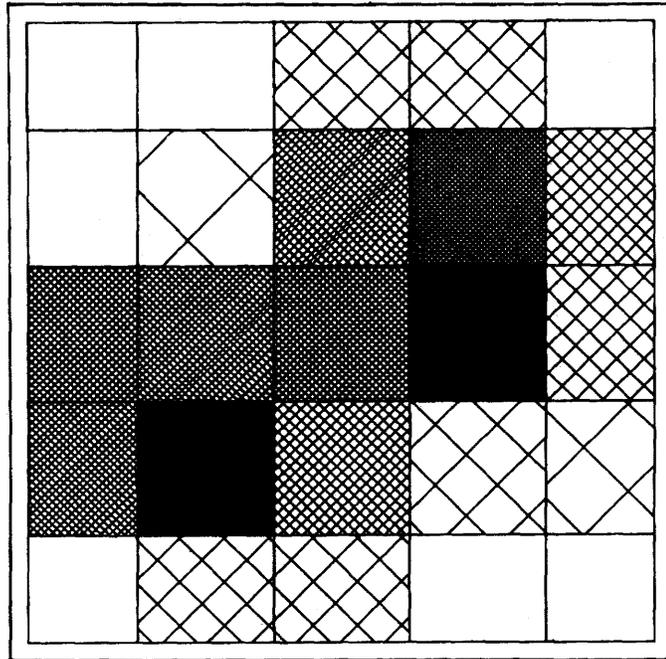
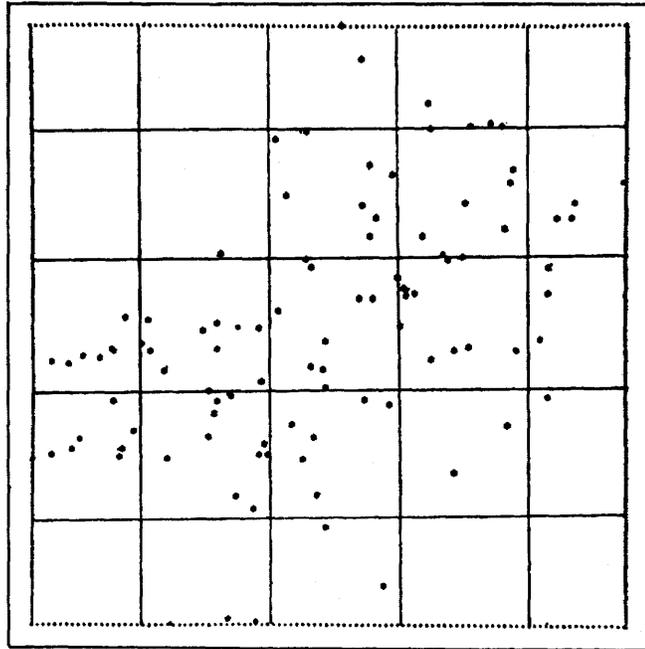
Booth, B., and Gétis, A. (1978), *Models of spatial Processes*, Cambridge: Cambridge University Press, 126—128.

A. 1970 Population Density by State^a



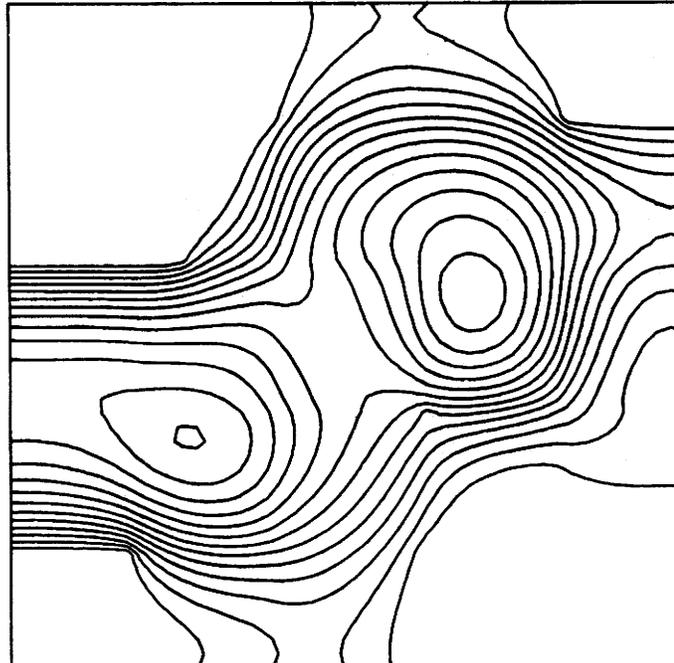
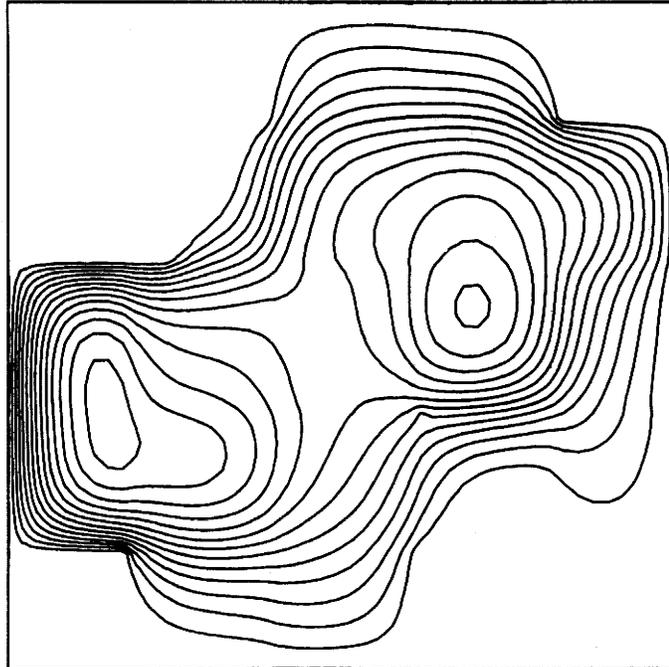
^a Computer-drawn bivariate histogram.

B. Test Example: Points and Packaging by Region^a



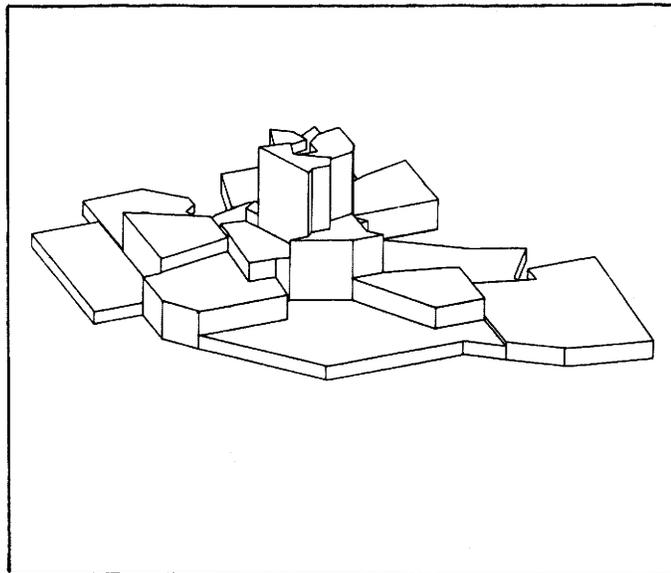
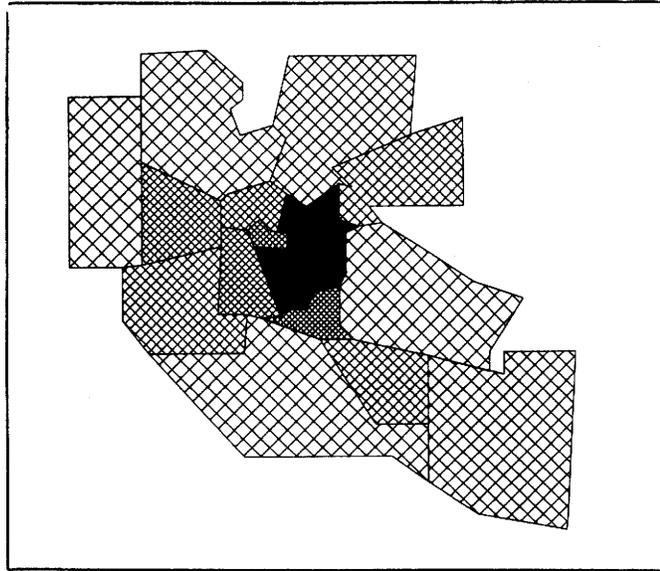
^a Top: Sample points taken from two overlapping bivariate normal distributions, with aggregation boundaries indicated. After Boneva, Kendall, and Stefanov (1971, Fig. 1, p. 45). Bottom: Density choropleths after aggregation into 25 rectangular regions and a border region.

*C. Isopycnics Computed From the Aggregated
Data of Figure B^a*

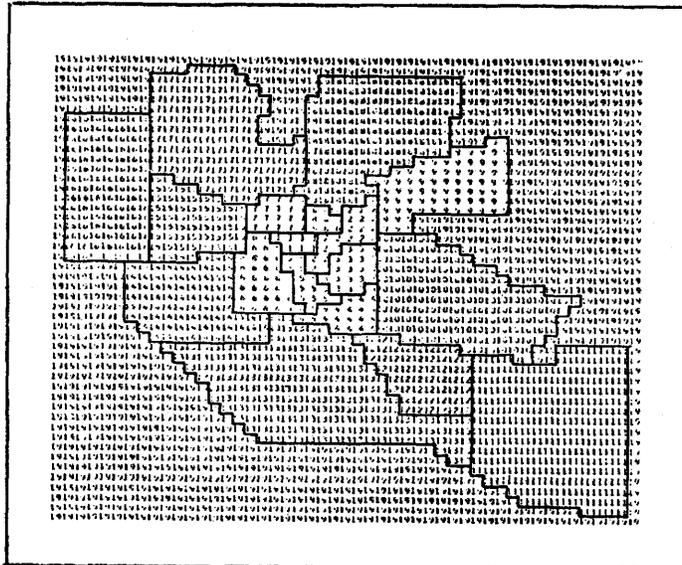


^a Contours shown are .01 (.01) .13. Top: Dirichlet condition with the border region set to zero. Bottom: Neumann condition using $\partial z/\partial n = 0$.

D. Population Densities in Ann Arbor Shown As Choropleths and As a Bivariate Histogram

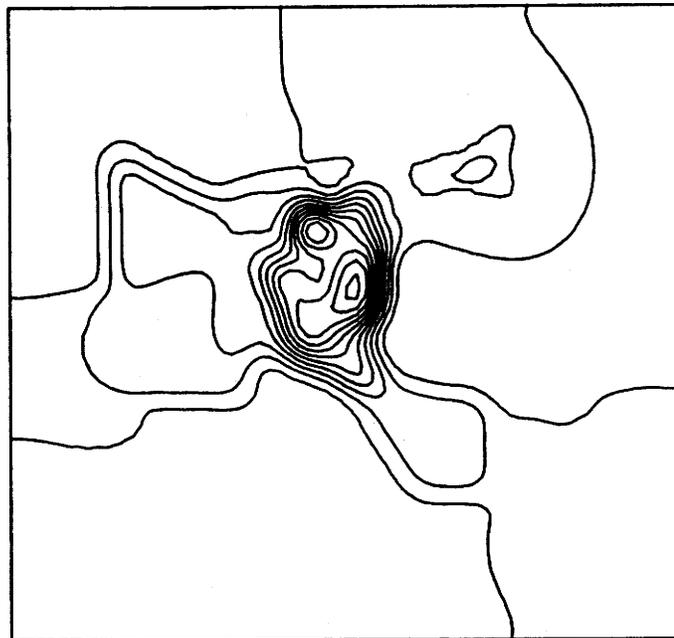
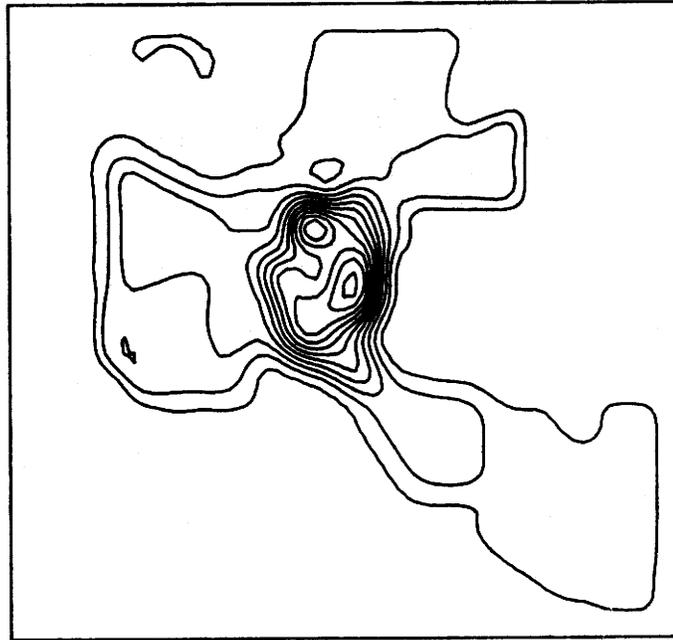


E. Approximation of Ann Arbor Census Tracts by a Lattice of Points^a



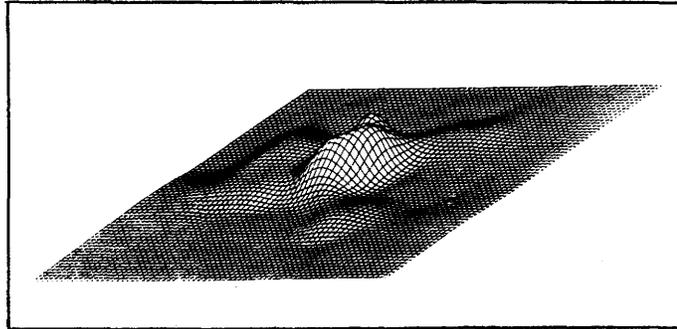
^a The actual mesh used is somewhat finer than here illustrated. The numbers identify the separate polygons.

F. Isodemopycnics Computed From the Data of Figure D^a



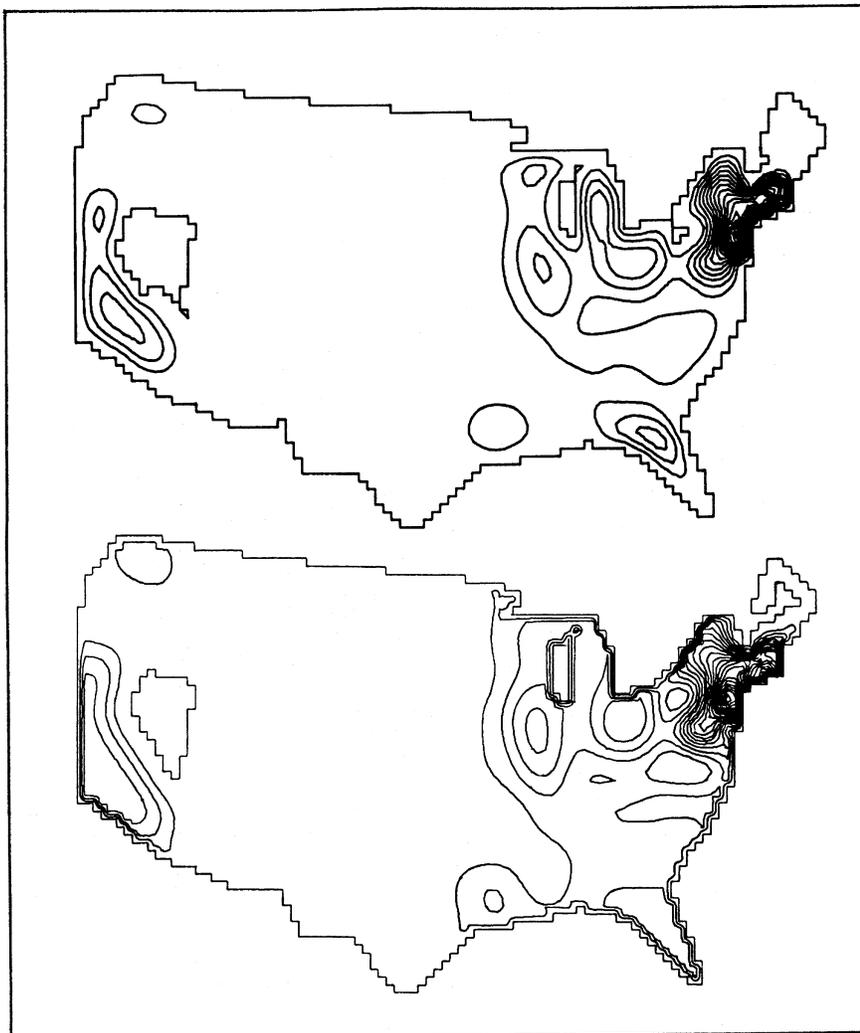
* The contours shown are 6(6)66 persons per hectare, approximately. Top: Using the Dirichlet condition with the exogenous area set to zero density. Bottom: Using the Neumann condition with $\partial z / \partial n = 0$ at the edge of the city.

*G. 1970 Population Density of Ann Arbor Based
on Census Tract Data^a*



^a Contours of Figure F, bottom, shown in an isometric rendering.

H. 1970 Population Density of the United States Based on State Data*



* Pseudophysical computation using the biharmonic equation as target. The contours are at 0(35)420 persons per km², approximately. Top: The Dirichlet case with the edge densities constrained to have the value zero. Bottom: The Neumann boundary condition $\partial z / \partial n = 0$ is used, with exogenous densities set to zero, resulting in a sharp drop along some of the edges.