Geo-Processing, 1(1980) 297—312
Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

AREAL INTERPOLATION: A VARIANT OF THE TRADITIONAL SPATIAL PROBLEM

MICHAEL F. GOODCHILD[1] and NINA SIU-NGAN LAM[1]

[1] Department of Geography, The University of Western Ontario, London, Ont. (Canada)

(Accepted 23 July 1980)

ABSTRACT

Goodchild, M.F. and Lam, N.S., 1980. Areal interpolation: a variant of the traditional spatial problem. Geo-Processing, 1:297-312.

A variant of conventional spatial interpolation is the problem of estimating aggregate statistics for one set of regions from comparable statistics for another set which do not necessarily respect the boundaries of the first. The paper discusses the properties of a technique of areal interpolation based on the use of areas of intersection as weights. The results are compared to those of several other methods, and illustrated with a case study.

INTRODUCTION

The term spatial interpolation generally refers to the problem of estimating the value of a variable z at some point (x, y) given known values of z at a number of data points, usually arranged randomly. As such it is inherent in contouring, and important in all spatially-oriented fields. A large number of approaches exist in the literature (see reviews by Crain, 1970; Lancaster and Salkaulkas, 1977; and Schut, 1976). The Kriging literature (see for example David, 1977) recognizes a variant of the problem, that of using point data to estimate integrals of the z function over arbitrary areas. A parallel example might be the use of determinations of population density at irregularly located points to estimate the total population within an arbitrary area. Another variant, the reverse of this, is Tobler's (1979a) pycnophylactic problem of estimating a population density function z at the nodes of a fine lattice from aggregate statistics on population for arbitrarily shaped regions; the pycnophylactic condition, which is imposed on the interpolated estimates, is that the integral of the density function within each data region be equal to the region's observed population.

In this paper we consider a further variation, which might be regarded as the logical extension of the previous two. Suppose population, or some other aggregate statistic, is known for a number of arbitrary divisions of an area, such as the

0165-2273/80/0000—0000/$02.50 © 1980 Elsevier Scientific Publishing Company

census tracts of a metropolitan region. A common problem is to obtain comparable estimates for a different set of regions which do not in general respect the boundaries of the first set. We will refer to this as the areal interpolation problem, and following Ford (1976) identify the two sets of regions as the source and target sets respectively. Let there be n source zones and m target zones, m generally not equal to n. Then the areal interpolation problem is to obtain $\underline{V}$, a column vector of length m whose elements are the target zone estimates, from the source zone statistics $\underline{U}$, a column vector of length n.

The areal interpolation problem occurs in a wide range of applications, in both analysis and planning, and particularly in urban areas. The districts used by census agencies, school boards, city government or voting systems rarely coincide, and so it is generally impossible to compare directly data collected or aggregated by different agencies. It is sometimes possible to avoid the problem by a re-aggregation from the individual level: for census data, it may be possible to recount or reaggregate the census to the target zones (see for example Statistics Canada, 1972). But in general this option, if available, is time consuming and expensive and often raises problems of confidentiality.

Two types of aggregate statistics must be identified. A statistic which is expected to take half the region's value in each half of a region is said to be spatially extensive; examples are population and gross income. A spatially intensive statistic is one which is expected to have the same value in each part of a region as in the whole; examples are average income or per cent male. To every spatially extensive statistic there corresponds a density function, which is obtained by dividing by area. Thus population density is obtained by dividing population by area, and yields population when integrated over area.

APPROACHES

A common approach to the problem is to reduce it to one of conventional spatial interpolation. A representative point is chosen for each source region, usually the centroid although this may not in fact lie inside the region. Each point is then assigned a representative value for the region, to be treated as a point estimate of a continuous function z. For spatially extensive data this value would be the region's value u_s divided by the region's area; for spatially intensive data u_s itself.

It is now possible to use conventional point interpolation procedures, in one of two ways. Values can be assigned directly from source centroids to target centroids using some weighting procedure based on the distances between them. Alternatively a fine lattice can be laid over the study area and the function z interpolated to each lattice node. The interpolated lattice values lying within each target zone are then summed in the case of extensive data to obtain an approximation to the integration of the continuous function z, or averaged for intensive data. The fineness of the lattice is clearly critical, particularly if there are small target zones, and there is very little understanding of how lattice size affects the accuracy of the estimates (Goodchild, 1980; Goodchild and Moy, 1977). It might be useful to distort the grid, to obtain a greater density of nodes in areas of small target zones, but this possibility does not seem to have been explored except in the literature on the numerical solution of partial differential equations, where it is a common practice. In both the direct and grid methods the spatial interpolation procedure should reflect prior expectations about the spatial variation of z; Tobler's (1979a) procedure for example maximizes smoothness on the interpolated surface consistent with the imposed boundary conditions and the pycnophylactic constraints.

The main focus of this paper is on an alternative approach which avoids the point interpolation step. Suppose the source and target regions are superimposed, the technical problem identified in the geographical data processing literature as polygon overlay (see for example White, 1978; Goodchild, 1978), and a matrix $\underline{A}$ defined whose elements a_{ts} are the areas of intersection or overlap between each target and each source. We now define a new matrix $\underline{W}$ by standardizing the elements of $\underline{A}$. For extensive data the matrix is standardized by column:

$$w_{ts} = a_{ts} / \sum_{t=1}^{m} a_{ts} \qquad (1)$$

so that w_{ts} gives the proportion of the area of source zone s located in target zone t. Then the target zone statistics $\underline{V}$ can be estimated by using $\underline{W}$ as a matrix of weights:

$$\underline{V} = \underline{W}\,\underline{U} \qquad (2)$$

For intensive data it is appropriate to standardize by row:

$$w_{ts} = a_{ts} / \sum_{s=1}^{n} a_{ts} \qquad (3)$$

to give the proportion of each target zone located in each source zone.

An example is shown in Figure 1. The light lines, the source zones, delimit 51 census tracts for the city of London, Ontario, as defined for the 1971 Census. The heavy lines are the boundaries of the 21 planning districts used by the city for population projection, planning of services, etc. Although the boundaries are sometimes coincident, there are major differences in the criteria used to design the two sets of zones. And though census data should play a major part in urban planning, it is impossible to use tract level data in any analysis of the planning districts.

As the method is intuitively simple and relies solely on the assumption of homogeneity within each source zone it is not surprising to find that it has been described in a number of disciplines in a widely scattered literature. Linsley,

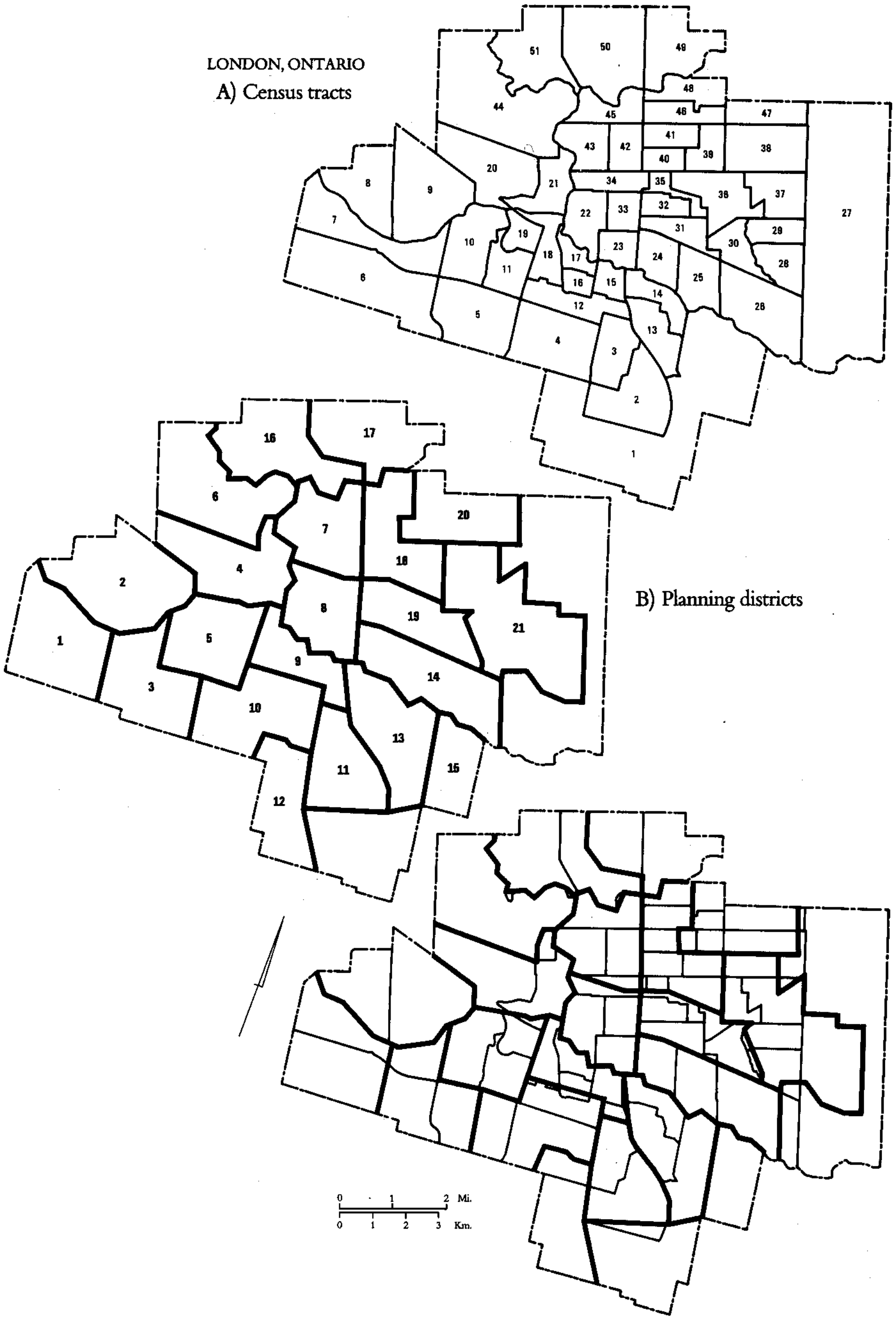

Figure 1. Overlay of census tracts (A) on planning districts (B).

Kohler and Paulhus (1958, p. 35) describe the use of area weighting to estimate the precipitation over a drainage basin from point records, a method derived ultimately from Thiessen (1911). Robinson, Lindberg and Brinkman (1961) used areal weights to convert county-based statistics to regular hexagons (see also Haggett, 1977, p. 288). The application of areal weights to density estimation in cartography is discussed by Robinson and Sale (1969, p. 106) based on Wright (1936). Markoff and Shapiro (1973) distinguish between extensive and intensive data (absolute figures and proportions in their terminology) and present equations similar to those above for estimating historical statistics for target zones from overlapping source zones. Crackel (1975) considered the problems which arise when either the source zones or the target zones, or both, do not completely partition the study area. In this paper we will assume that any part of the area not allocated to a source zone forms part of an additional dummy zone, and similarly for the target zones. One should note that this is necessary in the case of the example.

The next section of the paper is concerned with a discussion of the $\underline{W}$ matrix and its properties, and other aspects of this approach to areal interpolation. This is followed by a comparison of this and other techniques and an evaluation.

PROPERTIES OF THE $\underline{W}$ MATRIX

$\underline{W}$ is usually non-square and sparse. The number of nonzero entries reflects the degree of coincidence of boundaries of the source and target zones, and is minimum (the greater of m and n) when no source zone boundary ever crosses a target zone boundary. The maximum is mn since it is always possible to construct a set of target zones each of which overlaps each of a given set of source zones. In fact nonnegativity is the only general restriction on the elements of $\underline{A}$. For the example in Figure 1 133 of the 1071 entries are nonzero. It is possible, however, that many of the small nonzero entries in any $\underline{W}$ matrix are spurious. Stretches of boundary which are effectively coincident in reality, or identical with respect to the populations they contain, may diverge because of differences in legal definition, digitizing errors or cartographic generalization. This 'spurious polygon' or 'coastline weave' problem (Goodchild, 1978) will be reflected in trivial entries in the $\underline{W}$ matrix.

An interesting case arises when a subset of target zones and a subset of source zones overlap only with each other. The problem is said to be partitioned if a subset of target zones T and a subset of source zones S exist, such that $a_{ts} = o$ for all $t \in T$, $s \notin S$ and for all $t \notin T$, $s \in S$. It is assumed that the subsets are the smallest ones possible.

One property of this interpolation procedure with potential ramifications is the relative ease with which it can be reversed. Consider a spatial interpolation problem in which the values at one set of points, the target points, have been interpolated from another set, the source set. It is possible to reverse the problem by applying the same interpolation procedure to the target estimates to obtain revised values at the source points. Tobler (1979b) has argued that the difference between the original and revised source data values is a measure of the performance of the interpolation procedure, and has devised optimization methods based on such criteria. For most conventional methods the difference is unpredictable: distance-weighted averaging for example pays no attention to reversal properties. If trend surface analysis is used for interpolation, on the other hand, the differences are simply the residuals at the source data points.

Let $\underline{U}^*$ denote the revised source zone vector, estimated from the target zone vector $\underline{V}$ by reverse interpolation:

$$\underline{U}^* = \underline{B}\ \underline{V} \tag{4}$$

If the reverse is a straightforward application of the areal interpolation method, $\underline{B}$ will be an n by m matrix related to the transpose of $\underline{W}$. For extensive data

$$b_{st} = a_{ts} / \sum_{s=1}^{n} a_{ts} \tag{5}$$

Since the standardization is over a different set of elements, $\underline{B}$ will only be exactly the transpose of $\underline{W}$ when $\sum_{s=1}^{n} a_{ts} = \sum_{t=1}^{m} a_{ts}$ for all s and t, which implies that all source and target zones are of equal size, and therefore also that m = n.

Substituting for $\underline{V}$ we can now write

$$\underline{U}^* = \underline{B}\ \underline{W}\ \underline{U} \tag{6}$$

The product $\underline{B}\ \underline{W}$ is a square non-symmetric matrix. In the extensive form the columns of $\underline{B}$, $\underline{W}$ and $\underline{B}\ \underline{W}$ all sum to 1, and in the intensive form the same is true of the rows instead.

Although $\underline{B}\ \underline{W}$ determines the similarity between $\underline{U}^*$ and $\underline{U}$, it does so only in the context of some set of target zones, and not generally. For perfect reversibility of the interpolation, we require $\underline{U}^* = \underline{U}$, or

$$[\underline{B}\ \underline{W} - \underline{I}]\underline{U} = 0 \tag{7}$$

so that in general interpolation reverses perfectly only for certain sets of data. It occurs for all $\underline{U}$ if and only if $\underline{B}\ \underline{W} = \underline{I}$, which in turn requires exactly one nonzero entry in each row of $\underline{A}$, or spatially that each target zone be wholly within a single source zone.

Spatially we would expect that the condition $\underline{U}^* = \underline{U}$ also occurs for any $\underline{B}\ \underline{W}$ when $\underline{U}$ represents a constant density surface, or a constant spatially intensive statistic. Extensively, this implies that $u_s = \Delta \sum_t a_{ts}$, or a statistic proportional to source zone area, Δ being the constant density. Writing $\sigma_s = \sum_t a_{ts}$ for the area of each source zone and $\tau_t = \sum_s a_{ts}$ for the area of each target zone, the elements of $\underline{B}$ and $\underline{W}$ are

$$b_{st} = a_{ts}/\tau_t, \quad w_{ts} = a_{ts}/\sigma_s \tag{8}$$

and the elements of $\underline{B}\ \underline{W}$ are consequently

$$bw_{ij} = \sum_k a_{ki}\, a_{kj}/\tau_k\sigma_j \tag{9}$$

where bw_{ij} denotes an element of the product $\underline{B}\ \underline{W}$. Setting $u_j = \Delta\sigma_j$ it follows by manipulation that

$$u^*_i = \sum_j bw_{ij}\, u_j = \Delta\sigma_i \tag{10}$$

The condition $\underline{U}^* = \underline{U}$ therefore holds when $\underline{W}$ is defined for extensive data. For intensive data the appropriate condition is that all elements of $\underline{U}$ be equal to Δ.

When the problem can be partitioned, the source zones can always be permuted to give a pseudodiagonal $\underline{B}\ \underline{W}$ matrix: $bw_{ij} = 0$ wherever i and j are in different subsets. Any $\underline{U}$ now reverses perfectly when $u_j = \Delta_l\sigma_j$ for extensive data, or $u_j = \Delta_l$ for intensive data, where source zone j is a member of subset l and Δ_l is the density or intensive statistic for that subset. Spatially, this requires a constant density or constant intensive statistic within each partition of the problem.

In most cases, however, reversal is not perfect and $\underline{U}^* \neq \underline{U}$. For intensive data, the rows of $\underline{B}\ \underline{W}$ sum to 1, and thus $u^*_i = \sum_j bw_{ij}\, u_j$, $\sum_j bw_{ij} = 1$. It follows, since all bw_{ij} are nonnegative, that $\operatorname{Max}_i(u^*_i) \leqslant \operatorname{Max}_i(u_i)$ and $\operatorname{Min}_i(u^*_i) \geqslant \operatorname{Min}_i(u_i)$. In other words $\underline{B}\ \underline{W}$ can be regarded as a spatial averaging or smoothing operator. In the extensive case it is easy to show that $\underline{B}\ \underline{W}$ averages the densities u_i/σ_i.

Suppose now that $\underline{B}\ \underline{W}$ is applied repeatedly, to obtain $\underline{U}^{**}$, $\underline{U}^{***}$ and in the limit $\underline{\Theta}$. For intensive data $\theta_i = \sum_j u_j\sigma_j/\sum_j \sigma_j$, the area-weighted average of the original statistics, and for extensive, $\theta_i = \sigma_i \sum_j u_j/\sum_j \sigma_j$. In the latter case the total population, for example, has been redistributed with uniform density over all regions. However if the problem is partitioned, averaging occurs independently within each partition. These limits are consistent with the previous discussion, since we have already seen that any vector of the form $\underline{U} = \Delta\underline{\Theta}$ satisfies $\underline{U} = \underline{B}\ \underline{W}\ \underline{U}$.

In summary, then, reversability of the interpolation depends both on $\underline{U}$ and $\underline{B}\ \underline{W}$, and hence on the configuration of target zones. A measure such as the sum of squared differences $\sum_i (u^*_i - u_i)^2$ can be predicted only from the specific parameters of the problem, and is zero only for special configurations of the source and target zones, or for statistics with special spatial distributions. It is tempting to consider modifying the original $\underline{A}$ matrix, and thus $\underline{B}$ and $\underline{W}$, in order to minimize $\sum_i (u^*_i - u_i)^2$ but there are no obvious candidate parameters other than the specific terms of $\underline{A}$, and we know already that an $\underline{A}$ matrix can be found such that $\underline{U}^* = \underline{U}$.

EXAMPLE APPLICATION

In this section the area-weighted method described above is applied to the London Census Tract/Planning District problem, and its results compared to those of several other methods. The 1971 total population statistics were used as the Census Tract source zone vector $\underline{U}$, and exact tabulations for the target zones were obtained from Statistics Canada.

Table 1 shows the estimates of target zone populations obtained, and the actual values for comparison, and summary statistics are shown in Table 2. The original source zone populations are given in Table 3, together with the results of a single reversal, the largest errors in target zone populations occur where the assumptions of the method are least valid, where the population is least homogeneously distributed within the source zones. The target districts with the two largest errors are both large, suburban areas of comparatively low population where the corresponding source zones show rapid spatial variation in population density.

The same problem was subjected to a version of Tobler's (1979a) pycnophylactic interpolation. A 100 by 100 grid was laid over the area (using smaller grid cells gave no improvement in the results), and each grid cell allocated an initial

TABLE 1

Planning district population estimates

Planning district	Actual population	Overlay estimate	Error as per cent of estimate	Pycnophylactic estimate	Error as per cent of estimate
1	5407	4721.4	14.5	3167.7	70.7
2	11240	11869.8	- 5.3	12038.7	- 6.6
3	2745	5535.2	-50.4	6976.2	-60.7
4	9742	10355.6	- 5.9	10738.9	- 9.3
5	12764	13358.8	- 4.5	13299.8	- 4.0
6	6625	6655.8	- 0.5	5903.4	12.2
7	15527	15611.3	- 0.5	16531.8	- 6.1
8	17333	17362.0	- 0.2	19391.0	-10.6
9	17129	16185.8	5.8	15997.9	7.1
10	11190	9812.8	14.0	9843.3	13.7
11	5881	5675.8	3.6	5640.8	4.3
12	749	764.2	- 2.0	175.5	326.8
13	10554	10172.3	3.8	10225.8	3.2
14	17746	16956.6	4.7	16277.5	9.0
15	87	261.8	-66.8	125.3	-30.6
16	1906	2814.9	-32.3	3046.5	-37.4
17	6443	6371.8	1.1	5816.2	10.8
18	19352	16729.6	15.7	16054.4	20.5
19	15454	16264.9	- 5.0	15218.9	1.5
20	14317	12683.7	12.9	11530.8	24.2
21	18388	23070.9	-20.3	17446.9	5.4

TABLE 2

Summary statistics

	R^2	Mean absolute per cent error (unweighted)
Overlay	0.94	12.9
Tobler's Pycnophylactic	0.93	32.1
Distance-weighted average ($\beta = -0.15$)	0.26	61.1
Approximation I	0.84	18.4
Approximation II	0.88	19.5

TABLE 3

Original source zone population and reversal estimates

Tract	U	U*
1	795.0	1144.3
2	4785.0	4575.5
3	1205.0	1437.6
4	2165.0	2250.0
5	6470.0	6359.6
6	1900.0	1996.5
7	5010.0	4975.8
8	3865.0	3780.8
9	7830.0	7797.2
10	6925.0	6825.7
11	4190.0	4253.9
12	5370.0	5378.7
13	5580.0	5306.6
14	4780.0	4782.2
15	2140.0	2237.7
16	2860.0	2816.3
17	4905.0	4698.0
18	7360.0	7072.7
19	1520.0	1749.5
20	3420.0	3533.6
21	5980.0	5845.1
22	4540.0	4939.5
23	4750.0	4630.0
24	4990.0	4967.0
25	6505.0	6398.5
26	6405.0	6346.4
27	8235.0	8245.2
28	3565.0	3637.1
29	4565.0	4441.5
30	5870.0	5648.1
31	2630.0	3141.8
32	6010.0	5145.2
33	5170.0	5103.9

(cont'd)

TABLE 3 (cont'd)

Tract	U	U*
34	5075.0	4934.5
35	155.0	985.2
36	5415.0	5424.8
37	2815.0	2906.6
38	5505.0	5506.0
39	4715.0	4651.0
40	3825.0	3513.0
41	4185.0	4324.2
42	5035.0	5052.0
43	5480.0	5496.6
44	6770.0	6739.9
45	2555.0	2583.6
46	4515.0	4424.2
47	2505.0	2312.5
48	3880.0	3819.4
49	5170.0	5007.4
50	1820.0	1953.2
51	1525.0	1510.1

z value equal to the population density of the tract in which it fell. The smoothness condition was then imposed by replacing each cell's value by the mean of the values of the four neighbouring cells (the 4-neighbours or Rook's case neighbours), proceeding row by row from the top left corner. Cells outside the city limits were set to zero, and smoothing applied across the city limits to provide a suitable boundary condition.

After each smoothing step the pycnophylactic condition was reimposed in each polygon by adjusting each cell value by a correction factor. The process was continued until no further change occurred between two successive cycles. The results of applying pycnophylactic interpolation to the 51 census tracts to obtain planning district estimates are shown in Tables 1 and 2 for comparison with the overlay estimates. The errors show a somewhat similar pattern for both estimates (r = .30), but tend to be larger for the pycnophylactic.

Additional estimates were generated using a distance-weighted average. Each census tract was represented by a single population density value at the centroid (the centroid of one tract is actually outside its boundary) and the density interpolated at the centre of each cell of a 100 by 100 grid using the function

$$z(\underline{X}) = \frac{\sum_i z(\underline{Y}_i)\, 2^{|\underline{X} - \underline{Y}_i|/\beta}}{\sum_i 2^{|\underline{X} - \underline{Y}_i|/\beta}} \tag{11}$$

where $\underline{X}$ is the location of the centre of a cell

$Z(\underline{X})$ is the density at $\underline{X}$

$\underline{Y}_i$ is the location of the centroid of the ith tract

β is a parameter.

The function of β is to control the degree of smoothness through the relative weights given to nearby and distant tracts: β is the distance over which the assigned weight halves. The opportunity therefore exists to choose that β which minimizes the differences between planning district estimates and the known populations.

A value of $\beta = -0.15$ was found to both maximize R^2 and minimize mean absolute per cent error, and the results are shown in Table 2. This value of β is expressed in the length units of the coordinate system, and corresponds to roughly 2% of the E-W width of the city; in these units the smallest tract has an area of 0.174, which would give it a radius of 0.24 if it were circular.

The estimates from distance weighting are much poorer than the others. In part this is due to the absence of a suitable boundary condition: in any averaging process all interpolated values must lie in the range

$$\underset{i}{\mathrm{Min}}[Z(\underline{Y}_i)] \leqslant Z(\underline{X}) \leqslant \underset{i}{\mathrm{Max}}[Z(\underline{Y}_i)] \qquad (12)$$

It might be more appropriate to impose continuity of both Z and its first derivatives at the city limit, as the pycnophylactic method does, and this might give improved estimates. The quality of the estimates also depends on the grid cell size, but at this resolution the smallest tract was allocated 23 cells, suggesting that smaller cells would not improve the estimates substantially.

Of course these results give no indication of the general suitability of the three methods, but only an assessment for this particular problem. In general both overlay and Tobler's method have the advantage that they impose the pycnophylactic constraint. Tobler's approach imposes smoothness on the interpolated values, whereas overlay in effect allows discontinuities to exist in the density surface, by assuming homogeneity within each source zone. The former will therefore out perform the latter when smoothness is a real property of the data, and do worse when reality is closer to the discontinuous model. But they can both be expected to do substantially better than any distance-weighted averaging process.

APPROXIMATIONS TO THE OVERLAY METHOD

In this section we consider various approximations to the intersection approach. Overlay and the evaluation of the terms of $\underline{A}$ is computationally difficult and expensive, and so it is common practice to resort to approximation using a substitute matrix $\underline{A}^1$. Two common methods for extensive data are

I) $a^1_{ts} = 1 \text{ if } a_{ts} > a_{ks} \ \forall\ k \neq t, \text{ else } 0$ (13)

II) $a^1_{ts} = 1 \text{ if } a_{ts} > 0, \text{ else } 0$ (14)

For intensive data one simply considers the t^{th} row instead of the s^{th} column. For extensive data option I consists of assigning each source zone's population to the target zone with which it has the greatest intersection: for II, the source zone population is shared equally among all target zones with non-zero intersection.

These methods have considerable operational advantages; since the longest intersections, and the non-zero intersections can both be identified visually one avoids the need to compute overlays and measure areas, and multiply large numbers of terms. The quality of the approximation depends on the relative magnitudes of the non-zero a_{ts} terms in each column (or row, as appropriate). I gives good approximations if one term is dominant, while II is good if the variance is small. If the number of target zones is much greater than the number of source zones both perform very poorly.

Both methods were applied to the census tract/planning district problem. The application of I was strightforward and gave estimates (Table 2) which were comparable with, but not as good as the Tobler and overlay values (both I and II impose the pycnophylactic condition). Application of II is not as simple because of the trivial entries in the $\underline{A}$ matrix, which would have a serious effect on the estimates. The distribution of magnitudes of the non-zero entries in the $\underline{A}$ matrix suggested that .05 would be a suitable critical value: a total of 61 entries less than .05 were therefore rejected as trivial. The estimates found by applying II to the remaining non-zero entries are shown in Table 2. Although they are substantially better than those of distance-weighted averaging, they remain crude approximations to estimates based on $\underline{A}$.

THE HOMOGENEITY ASSUMPTION

The overlay method's accuracy depends solely on the degree of homogeneity of densities or intensive statistics within the source zones. The Census Tracts used in the case study are clearly not perfectly homogeneous with respect to population density, since the target zone estimates obtained from them are not perfect, but thus far no attempt has been made to evaluate homogeneity directly. Census Tracts are defined as aggregations of smaller units known as Enumeration Areas, with a population of approximately 500 each; there were 460 populated EA's within the city of London in 1971. In this section we consider reaggregating EA's under various alternative criteria, and the resulting effects on target zone estimates.

According to Statistics Canada (1972) Census Tracts are delineated according to the following criteria:

i) a population between 2,500 and 8,000 except for tracts in the central business district or institutional tracts which may have a lower population

ii) an area that is as homogeneous as possible in terms of economic status and social living conditions

iii) boundaries that follow permanent and easily recognized lines on the ground

iv) as much as possible a compact shape.

Homogeneity of population density is only weakly implied by (ii).

Let A_k denote the area of EA_k, and P_k its population. The aggregation of EA's into larger units is defined by the matrix with elements ε_{ks} such that $\varepsilon_{ks} = 1$ if EA_k is in aggregate s, else 0. The population density in aggregate s is thus

$$u_s = \sum_{k=1}^{N} \varepsilon_{ks} P_k / \sum_{k=1}^{N} \varepsilon_{ks} A_k \qquad (15)$$

where N is the number of EA's. Now assume that the target zones, the Planning Districts, are also aggregates of EA's, $\delta_{kt} = 1$ if k is in PD_t, else 0. The estimate of the PD's population based on overlay with the aggregates is

$$v_t = \sum_{p=1}^{N} \delta_{pt} A_p [\sum_{s=1}^{n} \varepsilon_{ps} (\sum_{k=1}^{N} \varepsilon_{ks} P_k / \sum_{k=1}^{N} \varepsilon_{ks} A_k)] \qquad (16)$$

In essence the aggregates constitute replacement census tracts, and if they can be designed with precisely homogeneous density the estimates v_t should correspond exactly to the true populations of the target zones, $\sum_{p=1}^{N} \delta_{pt} P_p$. The assignment matrix with elements ε_{ks} should be constrained by contiguity so as not to produce spatially fragmented aggregates.

One way of examining the effects of reaggregation would be to find a matrix of the ε_{ks} such that the residuals $|v_t - \sum_{p=1}^{N} \delta_{pt} P_p|$ are minimized. This is a mathematical programming problem but does not seem amenable to straightforward solution. Instead the effects were examined by grouping EA's using an adaptation of Ward's (1963) hierarchical grouping procedure. The problem is best seen as one of collapsing 460 regions into 51, one step at a time. In each step that pair of contiguous regions is merged which gives the minimum value of

$$D^2_{ab} = [P_a/A_a - P_b/A_b]^2 \qquad (17)$$

where a and b denote two groups of EA's and P and A their populations and areas respectively. The effect is to create 51 regions with a high degree of internal homogeneity with respect to population density.

The 51 regions or aggregates created by this process are markedly different from the 51 Census Tracts. EA's containing isolated apartment buildings or other multi-family structures remain as small regions, while the majority of the area of the city, being occupied by fairly uniform density single family housing, becomes a single large aggregate. Target zone estimates based on these new aggregates are very poor. Table 4 shows the agreement between the new estimates and the actual target zone populations.

TABLE 4

Estimates based on 51 new aggregates

Criterion		R^2	Mean absolute per cent error (unweighted)
Homogeneity		0.17	177.3
Homogeneity plus compactness	($\alpha = 10^6$	0.70	100.4
	($\alpha = 10^7$	0.94	19.8
	($\alpha = 10^8$	0.95	14.5
	($\alpha = 10^9$	0.96	14.3
Homogeneity with size constraint	($A_{max} = 6.0$	0.97	13.6
	($A_{max} = 5.1$	0.90	19.3

The estimates can be improved enormously by adding additional constraints on the aggregation process. A compactness criterion was added by extending the definition of D^2_{ab}

$$D^2_{ab} = [P_a/A_a - P_b/A_b]^2 + \alpha[X_a - X_b]^2 + \alpha[Y_a - Y_b]^2 \qquad (18)$$

where (X_a Y_a) are the coordinates of the centroid of region a, and α is a parameter. The higher the value of α, the greater the importance of the distance between the centroids of a and b in determining the sequence of grouping. The results for various values of α are shown in Table 4.

The results were also improved by the introduction of a size constraint; regions were joined only if the area of the resulting aggregate lay below a prescribed value. Again the results are shown in Table 4.

In summary, it appears that since homogeneity of population density is not an explicit design criterion for Census Tracts, reaggregation from the Enumeration Area level using homogeneity and compactness as criteria gives better target zone estimates, as shown by an increase in correlation from 0.94 to 0.97, although the mean absolute per cent error actually deteriorates slightly. It is clear, however, that since perfect homogeneity cannot be achieved, using it as the sole criterion results in very poor estimates because of the importance of source zone geometry.

CONCLUSIONS

Four of the methods discussed in this paper -- the Tobler approach and the three variants of overlay -- have a considerable advantage in that they impose a pycnophylactic or volume-preserving condition. Within this group one finds two extremes: overlay methods assume homogeneity of density within source regions and discontinuities between, while Tobler's imposes maximum smoothness of the interpolated surface.

There are of course arguments in favour of both. If one source zone is considered in isolation then homogeneity represents the most likely condition, but if the densities in neighbouring zones are known it seems reasonable to modify the estimates accordingly, by imposing some degree of continuity across the boundaries. On the other hand it is not clear that this degree of continuity should be maximal, and homogeneity is a common design criterion of source zones, particularly Census Tracts. Although the evidence in the case study seems to favour homogeneity and therefore overlay, the optimum method lies some where between the two extremes. Rather than maximal smoothness or maximal homogeneity, there is a need for an approach which imposes a degree of smoothness through some form of autocovariance function, perhaps estimated from the data (David, 1977). Meanwhile the choice between the two methods should be determined by the expected characteristics of the density surface at the scale of the source and target zones; the estimates will be as good as the implicit assumptions of the method chosen.

The results show that areal interpolation is capable of yielding good estimates of population at the Census Tract scale. Although both methods involve substantial computation, they should become increasingly attractive given the current levels of interest in geographical data processing problems.

REFERENCES

Crackel, T.J., 1975. The linkage of data describing overlapping geographical units -- a second iteration. Historical Methods Newsletter, 8:146-150.

Crain, I., 1970. Computer interpolation and contouring of two-dimensional data: a review. Geoexploration, 8:71-86.

David, M., 1977. Geostatistical Ore Reserve Estimation. Elsevier, New York.

Ford, L., 1976. Contour reaggregation: another way to integrate data in O.M. Anochie, editor, Computers, Local Government and Productivity 2, Papers, Thirteenth Annual Conference of the Urban and Regional Information Systems Association (URISA), Seattle.

Goodchild, M.F. and W.S. Moy, 1977. Estimation from grid data: the map as a stochastic process, in R.F. Tomlinson, editor, Proceedings of the Commission on Geographical Data Sensing and Processing, Moscow, 1976. Ottawa: International Geographical Union, Commission on Geographical Data Sensing and Processing, pp. 67-81.

Goodchild, M.F., 1978. Statistical aspects of the polygon overlay problem. Harvard Papers on Geographic Information Systems, 6.

Goodchild, M.F., 1980. A fractal approach to the accuracy of geographical measures. Mathematical Geology, 12:85-98.

Haggett, P., Cliff, A.D. and A. Frey, 1977. Locational Analysis in Human Geography. 2nd Edition, vol. 2. Edward Arnold, London.

Lancaster, P. and K. Salkaulkas, 1977. A Survey of Curve and Surface Fitting, University of Calgary.

Linsley, R.D. Jr., Kohler, M.A. and J.L.H. Paulhus, 1958. Hydrology for Engineers. McGraw-Hill, New York.

Markoff, J. and G. Shapiro, 1973. The linkage of data describing overlapping geographical units. Historical Methods Newsletter, 7:34-46.

Robinson, A.H., Lindberg, J.B. and L.W. Brinkman, 1961. A correlation and regression analysis applied to rural farm population densities. Annals of the Association of American Geographers, 51:211-221.

Robinson, A.H. and R.D. Sale, 1969. Elements of Cartography. 3rd Edn. Wiley, New York.

Schut, G., 1976. Review of interpolation methods for digital terrain models. Canadian Surveyor, 30:389-412.

Statistics Canada, 1972. GRDSR: Geographically Referenced Data Storage and Retrieval System. Ottawa.

Statistics Canada, 1972. Dictionary of the 1971 Census Terms. Ottawa.

Thiessen, A.H., 1911. Precipitation for large areas. Monthly Weather Review, 39:1082-1084.

Tobler, W.R., 1979a. Smooth pycnophylactic interpolation for geographical regions. Journal of the American Statistical Association, 74:519-530.

Tobler, W.R., 1979b. Lattice tuning. Geographical Analysis, 11:36-44.

Ward, J.H. Jr., 1963. Hierarchical grouping to optimize an objective function. Journal, American Statistical Association, 58:236-244.

White, D., 1978. A design for polygon overlay. Harvard Papers on Geographic Information Systems 6. Addison Wesley, Reading, Mass.

Wright, J.K., 1936. A method of mapping densities of population with Cape Cod as an example. Geographical Review, 26:103-110.