AN EVALUATION OF AREAL INTERPOLATION METHODS

Nina Siu-ngan Lam
Department of Geography
Ohio State University
Columbus, Ohio 43210

ABSTRACT

In solving the areal interpolation problem, the overlay and the pycnophylactic methods are believed to yield more accurate target zone estimates than the conventional approach since these two methods preserve original source zone values. The accuracy of the target zone estimates resulted from these two methods are found to be related to the number and the area of split source zones and the variation in values between neighbouring zones.

INTRODUCTION

The problem of obtaining data for a set of areal units (target zones), e.g. political district data, from another set of areal units (source zones), e.g. census tract data, that is, the areal interpolation problem, has become an increasing concern in the field of geographic and cartographic information processing. This problem often arises when two or more sets of areal data have to be included in a study.

Conventionally, this areal interpolation problem is solved by some isopleth mapping technique. A mesh of grids is first superimposed on the source zones and control points representing source zones are assigned. A point interpolation scheme is then applied to interpolate the value for each grid. Finally, the estimates of each grid point are averaged together within each target zone, yielding the final target zone estimates. Problems associated with this approach have been discussed (e.g., Hsu and Robinson, 1970). The most critical one is that the original source zone values are not preserved before aggregating into target zones. As a result, the final target zone estimates are less predictable and the errors are likely to be higher.

Two other methods for areal interpolation including map overlay and pycnophylactic interpolation have recently been suggested (Goodchild and Lam, 1980). A common characteristic of these two methods is that the original source zone values are preserved. Such a volume-preserving characteristic is considered highly desirable for areal interpolation since subsequent estimation of target zone values is less subject to error. The target zone estimates obtained by these two methods have been shown to be more accurate than those obtained by using the traditional approach (Lam, 1980). Different sets of assumptions and problems, however, are involved in these methods, which will affect the quality of the final target zone estimates. This paper examines the major factors affecting the reliability of these methods. The procedures and characteristics of these two methods are first briefly discussed. The error factors are then identified and their effects are modeled and experimented using fractal surfaces.

THE OVERLAY METHOD AND ITS ERROR FACTORS

Suppose there are n source zones and m target zones. The areal interpolation problem is to obtain the target zone estimates, represented by a column vector $\underline{V}$ of length m, from the source zone value $\underline{U}$, a column vector of length n. The overlay method simply starts by superimposing the target zones on the source zones, and a matrix $\underline{A}$ consisting of the area of each target zone in common with each source zone (a_{ts}) can be constructed. For data which are in the form of absolute figures or counts, such as population and total income, an estimate of target zone t is obtained by:

$$V_t = \sum_s U_s a_{ts}/\sigma_s \qquad (1)$$

where σ_s refers to the area of the source zone s. In matrix representation, $\underline{V} = \underline{W}\underline{U}$, where $\underline{W}$ is a weight matrix containing elements of a_{ts}/σ_s.

The estimation procedure differs slightly for density data (e.g., population density) and ratio data (e.g., percent of male). The estimation formulae for these types of data can be found in Lam (1982).

It has also been shown before that for every target zone estimate, there is a theoretical maximum error range (Lam, 1982). For example, for data in the form of absolute figures, the maximum error range is simply the sum of values of all split source zones involved. For density data, the error range is equal to:

$$\sum_k U_k \, a_{tk}/\sigma_k \qquad (2)$$

where k refers to the split source zones involved in target zone t. T_t is the area of target zone t. In reality, the estimation error for every target zone estimate lies within the theoretical maximum and depends on a number of factors.

The most important error factor is that the overlay method assumes homogeneous source zones. This error factor is easily perceived and its effect on areal interpolation has been demonstrated (Ford, 1977; Goodchild and Lam, 1980). Unfortunately, the degree of homogeneity within source zones is always unknown since the source zones available are supposed to be the finest resolution one can obtain. This implies that the errors involved in the overlay estimates cannot be determined unless some known parameters which may serve indirectly as indicators of homogeneity are determined first.

Very few studies have been focused on the homogeneity of the areal units and its relationships with other factors despite its importance and frequent uses in many fields. Coulson (1978) was among the few to suggest that the size and the shape of the areal units are the two major factors in estimating the potential for variation within these units; the smaller and the more compact in shape the areal units, the lower the potential for variation within these units. These two factors have recently been examined by MacEachren (1982) and were further illustrated by his experiment. However, an independent study conducted by this author (Lam, 1982) did not support their findings. Thus the effect of size and shape on the homogeneity of areal units remains questionable. It is necessary to

examine other factors.

Since the target zone estimates are solely dependent on the weights derived from the area of intersection, a close look of the form of the weight matrix $\underline{W}$ will be useful. For data in the form of absolute figures, perfect results can be obtained in two extreme cases; where the source zones and the target zones having the same size and shape (Fig. 1a), and where the target zones being simply aggregates of source zones (Fig. 1b). In these two cases, the non-zero entries in each column are all equal to 1. In other words, there is no split source zones involved in each target zone. Any deviation from these two forms of matrices will result in certain estimation errors for target zones. The general case would be when the source zones and the target zones have different sizes and shapes, so that the weight matrix has more than one non-zero entries in each row and column, with most of them smaller than 1 (Fig. 1c). It is therefore expected that the accuracy of the estimates will be influenced by the number of non-zero entries which are smaller than 1, i.e., the number of split source zones involved, and more split zones may imply more possible sources of error. Similarly, it is expected that the proportion of area of a target zone occupied by the split source zones will also affect the quality of the target zone estimates.

It should be noted that the above relationship will be biased by the "coast-line weave" or trivial polygon problem (Tomlinson, 1972). This problem arises when the same boundaries in reality diverge slightly because of source map errors, digitization errors, or difference in zone definitions. In such cases, most of the non-zero entries in the weight matrix are trivial and will not reflect the target zone errors that are potentially involved.

In addition to the number and the area of split source zones, variation in values between a particular source zone and its neighbors may also serve indirectly as indicators of homogeneity. This corresponds to one of Ford's (1977) findings in his study on areal interpolation using the traditional approach. It might be reasonable to expect that higher variation between neighboring zones may imply a rougher underlying surface, and the zones delineated from this surface will likely be less homogeneous. For the present problem, since only the split source zones will contribute to the error, the mean absolute difference between each source zone and its neighbors, d_t, will be used in testing the model. In short,

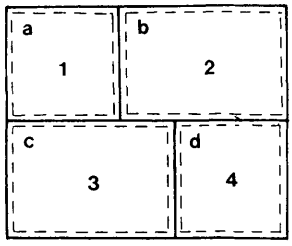
$$e_t = f(n_t, a_t, d_t) \tag{3}$$

where e_t denotes the error of the target zone estimate as represented by the difference between the estimated value; n_t, a_t are the number and the area of split source zones. Since there is very little knowledge about the manner in which the accuracy of the estimates are affected by these factors, a linear form of the above model will be examined at this preliminary stage.

THE PYCNOPHYLACTIC METHOD AND ITS ERROR FACTORS

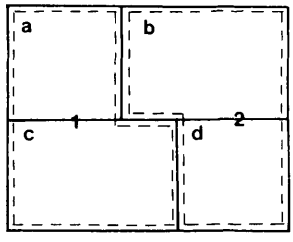
The pycnophylactic interpolation method was first suggested by Tobler (1979) for isopleth mapping and has recently been applied to areal interpolation (Goodchild and Lam, 1980). The method also utilizes an overlaid mesh of grids on the source zones. The grid values are esti-

a. Special case 1



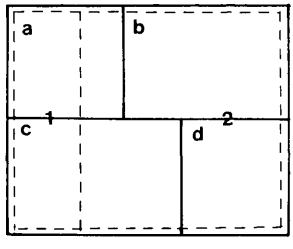
Target \ Source	a	b	c	d
1	1	-	-	-
2	-	1	-	-
3	-	-	1	-
4	-	-	-	1

b. Special case 2



Target \ Source	a	b	c	d
1	1	-	1	-
2	-	1	-	1

c. General case



Target \ Source	a	b	c	d
1	.6	-	.5	-
2	.4	1	.5	1

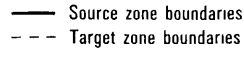


Figure 1. Configurations of source and target zones.

mated according to two criteria. First, the resultant surface is required to be smooth using some governing density function. One type of smoothing condition can be obtained by requiring the value of any grid cell be close to the average of its four neighbors. Second, the sum of all the grid values within a source zone is required to be equal to the original source zone value, i.e., the pycnophylactic or volume-preserving condition.

In short, the interpolation procedure begins by assigning the mean density to each grid cell, and then modifies this by a slight amount to bring it closer to the value required by the governing density function. The volume-preserving condition is then enforced by either incrementing or decrementing all of the densities within individual zones after each computation. Since the assignment of values for cells outside the study area will affect the measure of smoothness near the edge and consequently inward, the selection of a boundary condition should be careful. Different boundary conditions can be used. For example, a zero density may be assigned to outside areas when dealing with a study area bounded by water. The above procedure can be used for data in the form of absolute figures and for density data. For ratio data, similar to the overlay method, two separate procedures for the numerator and the denominator are required.

Compared with the overlay method, the pycnophylactic method represents a conceptual improvement since the effect of neighboring source zones have been taken into account, and secondly, homogeneity within zones is not required. The overlay method assumes a discrete surface with breaks along the source zone boundaries and the pycnophylactic method assumes a smooth surface as designated by the smooth density function. It is clear that the smooth density function and the boundary condition imposed are again only hypotheses about the surface and may introduce estimation errors for the target zone values.

Similar to the overlay method, the theoretical maximum error range for every target zone estimate, in the case of absolute-figure data, is the sum of values of all split source zones involved. For density data, the error range is the same as equation (2). This is in fact a fundamental characteristic of the volume-preserving areal interpolation methods. Again, it is expected that the number and the area of split source zones will affect the accuracy of the target zone estimates, though the manner and the magnitude of their effects may be different from those on the overlay method. For the factor of the variation between a source zone and its neighbors, since the pycnophylactic method assumes a smooth surface, it is expected that higher variation imply a rougher surface and as a result a higher chance or error. Hence, the linear model used for overlay may also be used for pycnophylactic in evaluating the reliability of these methods.

EVALUATION AND DISCUSSION

Evaluation of the above model includes two steps. First of all, computer-generated fractal surfaces of specific demensionalities were used. These surfaces are believed to be very useful for simulating the surface of the Earth (Mark, 1979). Higher dimensionality ($D>2.5$) means a rougher surface and lower dimensionality ($D<2.5$) results in a smoother surface. A dimensionality of 2.3 is found to correspond to most real-world surfaces (Mandelbrot, 1977). Four surfaces of dimensionality $D=2.1, 2.3, 2.7, 2.9$ were generated in the form of 30x30 grids using Goodchild's algorithm (1980) (Fig. 2a). They were then partitioned

into rectangles of different sizes to represent source and target zones. Figure 2b is an example of the hypothetical source and target zones. In this case the true values for target zones are known. The absolute difference between the estimated and the true target zone values as a percent of the estimated value (APE), the number and the area of split source zones, and the area-weighted mean absolute difference between each split source zone and its neighbors were calculated accordingly.

Initial stepwise regression of these variables show that there is only a moderate to weak relationship between the error as represented by APE and the three independent variables, and this relationship is unstable, with multiple R's ranging from 0.24 to 0.91 (Table 1). Secondly, surface complexity seems to have little effect on the behavior of the model since R's do not vary significantly among the four surfaces. This is mainly due to the fact that the surfaces were further partitioned in a random manner and the effect of surface complexity has thus been reduced.

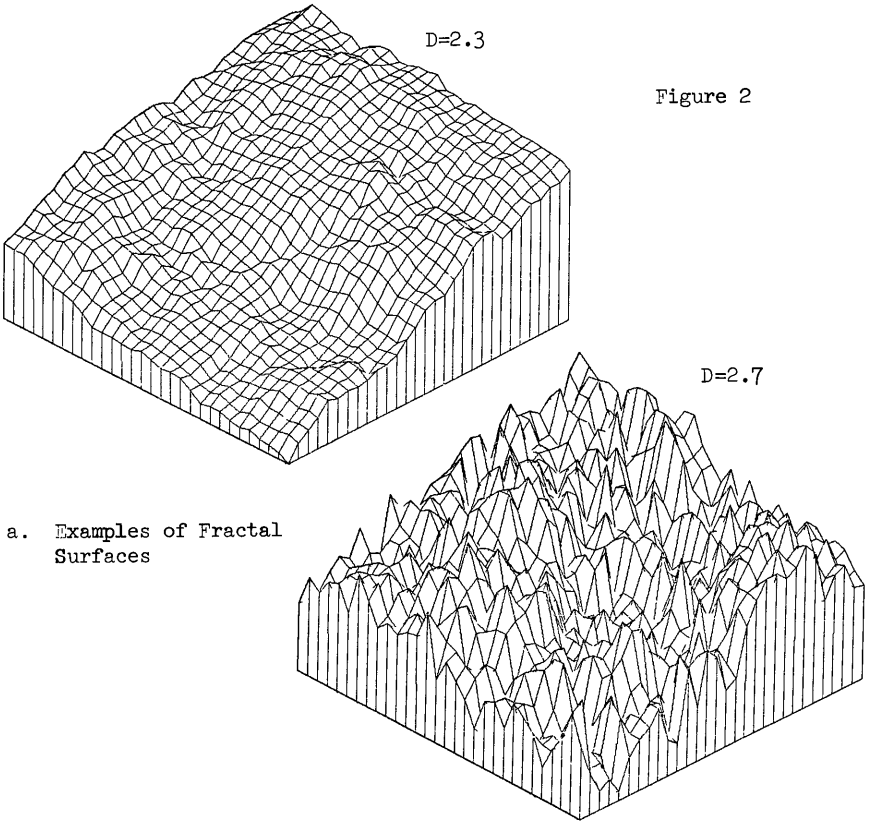
Table 1 : Summary Statistics

Surfaces	Size		#source/	Multiple R's			
	Source Zone	Target Zone	#target	Overlay	Pycno.*	Overlay	Pycno.**
	5x5	7x7	36/25	0.54	0.27	0.49	0.44
D=2.1	4x4	7x7	64/25	0.49	0.91	0.99	0.95
	5x4	7x5	48/30	0.47	0.65	0.62	0.35
	5x5	7x7	36/25	0.51	0.44	0.44	0.38
D=2.3	4x4	7x7	64/25	0.23	0.90	0.99	0.95
	5x4	7x5	48/30	0.35	0.24	0.74	0.57
	5x5	7x7	36/25	0.32	0.24	0.54	0.49
D=2.7	4x4	7x7	64/25	0.44	0.91	0.99	0.94
	5x4	7x5	48/30	0.57	0.47	0.60	0.50
	5x5	7x7	36/25	0.40	0.23	0.54	0.50
D=2.9	4x4	7x7	64/25	0.56	0.92	0.99	0.94
	5x4	7x5	48/30	0.38	0.60	0.70	0.58

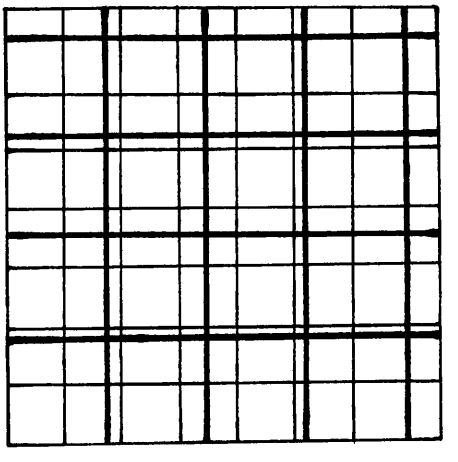
* Multiple regression using APE as dependent variable.
** Multiple regression using adjusted APE as dependent variable.

A third finding of this initial analysis is that the model performs better for overlay than for pycnophylactic, with exceptions, occur in the cases of 64/25 #source/#target zones. In these cases, R's are unusually high (>.90). A close examination of the values in these cases indicates that most source zones along the border are unable to maintain the original values after 100 iterations in the pycnophylactic interpolation process. This is largely due to the fact that the 30x30 grid mesh used is not fine enough for maintaining both the volume-preserving and the smoothing conditions. As a result, the estimation errors for the target zones along the border are higher, and coincidently, the number of split source zones are also smaller for these target zones, yielding higher R's than other cases. The failure to preserve the original source zone values may also contribute in part to the poorer performance of the model for the pycnophylactic method.

Figure 2



a. Examples of Fractal Surfaces



b. Hypothetical Source and Target zones

The moderate to low R's resulted and the instability of the model may be due to a number of factors, including for example incorrect form of the model, inappropriate definitions of the dependent and the predicting variables, and exclusion of some other factors. Therefore in a second step, an adjusted APE was used. The adjusted APE was derived by dividing the original source and target zone values by their areas first before interpolation. Stepwise regression were rerun using the adjusted APE as dependent variable. The resulting multiple R's increase substantially in all cases and for both the overlay and the pycnophylactic methods, though the relationship still remains fairly unstable.

In short, the above analysis has demonstrated that there is only a moderate and an unstable relationship between the accuracy of the target zone estimates and the error factors as defined. The second analysis suggests that further improvement on the model could be made in several ways. First of all, the definitions of the variables could be modified. The higher R's obtained from using the adjusted APE is one example. Secondly, different forms of the model may also be used instead of the linear one. For example, although a higher number of split source zones may likely include a larger amount of error, it is also expected that if these split source zones occupy only a small portion of the target zone area, then the error will likely be smaller. So these two factors could compensate each other and in mathematical form they should multiply each other instead of as two independent variables. In addition, the effect of the number of split source zones on the estimation error may also be reduced if most of the split zones are homogeneous. Again, these two factors could compensate each other by multiplication. Finally, different sizes and shapes of source and target zones could be used to encompass a wider range of values in the variables of the number and the area of split source zones.

CONCLUSION

This paper has illustrated that for every target zone estimate resulted from using the overlay and the pycnophylactic methods, there is a theoretical maximum error range, which is a major characteristic of the volume-preserving areal interpolation methods. In estimating the error within the range, several factors are suggested. They include the number and the area of split source zones and the variation in values of neighboring zones. The initial analysis has indicated that a moderate to weak linear relationship exists between the estimation error and the selected factors. However, redefinition of the error variable has improved the relationship substantially. It is suggested, therefore, that improvement of the model could be made in future studies by modifying the form of the model and the definitions of the variables.

REFERENCES

Coulson, M.R.C., 1978, "Potential for Variation: A Concept for Measuring the Significance of Variation in Size and Shape of Areal Units," Geografiska Annaler, 60B, 48-64.

Ford, L., 1976, "Contour Reaggregation: Another Way to Integrate Data," Papers, Thirteenth Annual URISA Conference, 11, 528-575.

Goodchild, M.F., 1980, "A Fractal Approach to the Estimation of Geographical Measures," Mathematical Geology, 12, 2, 85-98.

Goodchild, M.F., and Lam, N.S., 1980, "Areal Interpolation: A Variant of Traditional Spatial Problems," Geo-Processing, 1, 297-213.

Hsu, M.L. and Robinson, A.H., 1970, The Fidelity of Isopleth Maps - An Experimental Study, University of Minnesota Press, Minnesota.

Lam. N.S., 1980, Methods and Problems of Areal Interpolation, Ph.D. Dissertation, University of Western Ontario, London, Ontario.

Lam, N.S., 1982, "Areal Interpolation Using Map Overlay," Modeling and Simulation (Forthcoming).

MacEachren, A.M., 1982, "Thematic Map Accuracy: The Influence of Enumeration Unit Size and Compactness," Technical Papers of the American Congress on Surveying and Mapping, Denver.

Mandelbrot, B.B., 1977, Fractals - Form, Chance and Dimension, Freeman, San Francisco.

Mark, D.M., 1979, "Fractals - Form, Chance and Dimension: A Book Review," Geoprocessing, 1, 202-204.

Tobler, W.R., 1979, "Smooth Pycnophylactic Interpolation for Geographic Regions," Journal of the American Statistical Association, 74, 367, 519-536.

Tomlinson, R.F., et.al., edited, 1972, Geographical Data Handling, Vol. 1 & 2, IGU Commission on Geographical Data Sensing and Processing, Ottawa.