Keith C. Clarke
Hunter College
New York, NY 10021

A Comparative Analysis of Polygon to Raster Interpolation Methods

Interpolation methods which convert continuous spatial data, that have been sampled by polygons, to pixel-based data, for use as a GIS data layer, are compared.

Abstract: *Increasingly, data collected by irregular polygonal geographic regions are being converted to a grid-cell or pixel-based format for use as additional data layers within a Geographic Information System. This study used hypothetical data to examine the error involved in the conversion of polygonal data to a pixel-based format. Error was found to be related to the complexity of the underlying surface and the characteristics of the polygons. Conclusions were that the frequently-used method of direct overlay performs surprisingly well in general, but that careful choice of conversion method can minimize the associated gridding error. Implications for the choice of interpolation model were considered.*

Introduction

In remote sensing, data are usually collected in the context of a spatial grid, with single or multiple reflectance values corresponding to a pixel which is square or rectangular. Corrected for atmospheric, geometric, and flight anomalties, these data are used in Geographic Information Systems (GISs) after having been resampled into a uniform square grid based upon some appropriate map projection. In some cases, additional data layers are added to the remotely-sensed data, using the same grid-cell or pixel basis. Perhaps the most frequently-used data layer is topographic elevations, but increasingly other thematic layers, such as land ownership, land-use, roads, and administrative areas are used. In this way, the remotely-sensed data become a data input mechanism for GISs, giving all the advantages of timeliness, cost-efficiency, and access. The role of remote-sensing in this context has been considered at length (Estes, 1982) and is widely interpreted as an important future direction for the discipline. Similarly, cartographers have realized the potential of integrating remotely-sensed data into GISs for the purposes of compiling, updating, and evaluating maps.

The integration of remotely-sensed data into GISs, however, introduces a conflict of data structures. Pixel-based data consist of spatial arrays of spectral measurements. The pixel-based data layers have often been classified or processed so that the data values within a pixel correspond to land-use, geology, tree type, or some other land surface characteristic.

On the other hand, data from sources such as maps and aerial photos are usually vector-based, and consist of strings of eastings and northings defining line segments and polygons. Typical data structures are points, chains (lines made up of points), networks, and polygons. Points may be isolated sample sites or the centers of areas. Chains may be roads or rivers, and are often linked together by indices so that they can be converted to polygons. Networks consist of points (nodes) and links (chains) which usually have associated connections, linkages, and flows (Cox and Rhind, 1978). Two types of polygon systems are used. First, tesselations of geometric shapes such as triangles may be used (Peucker *et al.*, 1979). The second type is the irregular polygon or resolution element. This unit, the "resel," is usually defined by sets of points delineating the boundaries of regions. These regions combine to form a non-overlapping division of space, and may contain holes. Good examples of resels are the polygons drawn on the U.S. Geological Survey (USGS) 1 to 250,000 series land-use maps.

Photogrammetric Engineering and Remote Sensing,
Vol. 51, No. 5, May 1985, pp. 575-582

0099-1112/85/5105-0575$02.25/0

Of particular concern here are the resels used to take measurements of continuous spatial attributes, those attributes which need to be integrated with pixel-based data. These attributes may be human, such as population density and dollar expenditures, or natural, such as climatic variables, vegetation characteristics, or hazard potential.

Within GISs, the problem becomes the conversion of resel-based to pixel-based data. Once an appropriate resolution has been chosen, the conversion of the actual line information such as polygon boundaries is comparatively simple (Peuquet, 1981). Several GISs are able to cope with this transformation, among them the IBIS system (Zobrist, 1977). When areal data are converted, however, the problem is less simple. Simply gridding the polygon produces a resultant surface which contains sharp discontinuities, a condition which is not compatible with continuous spatial data. In all cases, the discontinuous data can be represented either cartographically (Jenks, 1963) or mathematically (Nordbeck and Rystedt, 1970) as continuous spatial functions. This implies that when data are sampled using a set of irregular polygons, such as census tracts or school districts, data values are averaged spatially over the polygon to produce a single value per polygon. This type of data, with a single data value (usually a ratio such as population density) and a corresponding bounded area on the map, is called choroplethic in cartography. An important fact to realize is that usually the choroplethic data are the only sample available. This implies that the real distribution is unknown, and has to be interpolated from the sample; and that any given interpolation of the real surface is an estimate with an associated error distribution.

There are three ways to convert choroplethic data to a pixel format. First, we can make the "resolution" of the irregular polygons high in relation to the pixel resolution. This means that we are simply changing the shape of the tesselation from irregular to regular, without much change in the data. The number of polygons will be similar or more than the number of pixels, resulting in generalization. This sort of interpolation has been used in statistical and meteorological work since at least 1961 (Robinson *et al.*, 1961), but suffers from the disadvantage that spatial variation at the order of magnitude below the pixel resolution is lost, producing highly generalized GIS data layers with a poor match between the levels of spatial variation for different types of data.

The second conversion is by far the most commonly used and might be called the direct overlay method, where pixels are assigned the choropleth value for the polygon inside of which they fall. Some decision rules are necessary at polygon boundaries, usually involving either the allocation of the choropleth value at the cell center or allocation of the value associated with the majority of the cell's area. Direct overlay has the advantage of being comparatively simple to perform, and is therefore used in many GISs. The level of error associated with the method has been the subject of several publications (e.g., Muller (1977)), and the technique has been tested against others in the context of converting between sets of overlapping polygons (Goodchild and Lam, 1980).

Direct overlay, however, incorporates an unrealistic assumption about the general distribution of the attributes in space. In cartography this assumption is known as the choropleth assumption, and states that all surface variance (i.e., variance summed by pixel for the whole image or map) is zero within individual polygons. This is in direct contradiction to one of the most fundamental geographic properties, the neighborhood relationship. This relationship may be stated verbally as "near things are related to each other," or mathematically as the autocorrelation function.

The third polygon to raster conversion attempts to incorporate the neighborhood relationship into the resulting pixel data. This is done using a model which imposes a given autocorrelation function onto the data. Different models form the basis for a variety of algorithms which are frequently used to interpolate data collected for irregular polygons to a regular grid.

Model-Based Polygon to Raster Conversion

Two types of models are usually used to convert polygonal data to a grid. These are point interpolation models and areal interpolation models.

point interpolation

Point interpolation models choose a point to represent the polygon and then interpolate pixel values from the irregularly spaced points which result. A number of parameters can be varied using this model, including the search method for finding nearby points, the number of points to be included in the computations, and the formula used to weight point values with respect to distance from the pixel. Some models involve processing the surface as a whole, such as polynomial trend surfaces, trigonometric series, or kriging. Other models are applied over the local neighborhood and iterate over the surface, such as inverse-distance weighting and trend projection methods. Reviews of the point interpolation methods can be found in Walters (1969), Harbaugh and Merriam (1969), Crain (1970), Shepard (1968), and Rhind (1971), which date from the period of research activity on these techniques. Some methods use sequences of models in multiple passes.

areal interpolation

A second approach to data conversion is to operate on the overlay matrix itself to erode the discontinuities by filtering. Hsu (1975) used variable

size filters to smooth the overlay matrix, where the size of the filter approximated the size of the polygon. Tobler (1979) used a technique called pycnophylactic interpolation. This method involved the iterative smoothing of the overlay matrix to conform to a set of smoothness criteria and boundary constraints. Most important, after each complete pass of the smoothing filter, values were adjusted so that choropleth means were retained within polygons. This implies that pycnophylactic interpolation smooths the overall image, reducing variance for the whole map, but adding variance within polygons.

The addition and subtraction of variance at different levels of resolution suggests that these techniques are simultaneously performing generalization and enhancement. Because no real estimates of surface or polygon variance are available in most cases, the model supplies the variance. This is done in two ways. Emphatic enhancement takes place when information about the autocorrelation function gained from parts of the image with high resolution polygons is used to "fill-in" variation in part of the image where the polygons are large. Synthetic enhancement, on the other hand, uses a model to add variation within the polygons. A technique based on synthetic enhancement is Fourier synthesis (Clarke, 1984), which allows the user to target how much total image variation is required, and then distributes it among the polygons based on the significant harmonics in the surface.

Three point interpolation and two areal interpolation methods were selected for comparative testing. Three of these operated on point data, in this case polygon centroids as given by the mean x and y values of the polygon definition. Two of the methods involved inverse distance weighting. The actual interpolation procedures operated by (1) selecting the five nearest polygon centroids for each pixel, (2) computing their distances from the pixel center, (3) computing the inverse of this distance to the power n for each value, (4) multiplying each point data value by the result, producing the weighted values, (5) summing the weighted values and assigning the total to the current pixel, and (6) repeating for all of the pixels. The inverse-squared ($n = 2$) and the inverse-fourth ($n = 4$) models were used.

The third technique, that of trend projection, operated in two phases. Phase one involved a parse of all centroids, during which each data value was replaced by an estimate from a least-squares fit of a linear trend surface to the centroid's eight nearest neighbors. The adjusted data were then used as input to phase two, which was identical to the method described above with $n = 2$, as recommended by Sampson (1978).

The two remaining were pycnophylactic interpolation (Tobler, 1979) and spatial smoothing, adapted from Hsu (1975). In the former case, the Neuman condition was used for the edge while the Laplace equation was used for overall surface smoothness. For surface smoothing, a single Hanning filter of a size equal to the area of the mean polygon was used. This represents a slight change from Hsu's method, but was conceptually and computationally simpler.

The Test

Five mathematical equations were used to generate grid values for a 50 by 50 grid. The surfaces chosen were not unlike many continuous distributions associated with human and natural geographic features. These were a linear trend surface, a cubic polynomial surface, an exponential peak, a mixture of harmonics in x and y, and the same surface after the application of a random filter (Figure 1). Examples of each surface may be mean January temperature, pollution from a point source, population density in a city, depths of glacial till associated with a drumlin field, and the depth of the till in the same drumlins after substantial erosion.

Six sets of polygons were used in the test (Figure 2). These sets of polygons mixed the size, shape, number of vertices, orientation, and spatial varia-

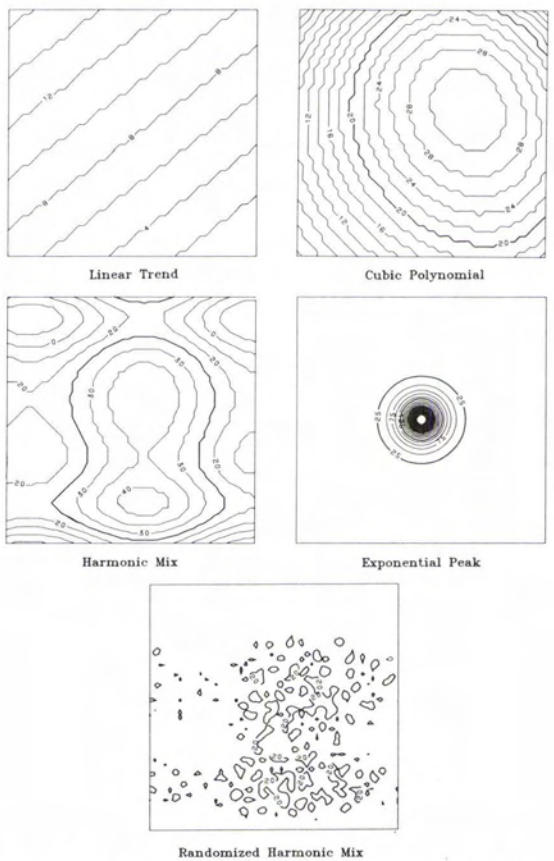

FIG. 1. Hypothetically generated surfaces used in the test.

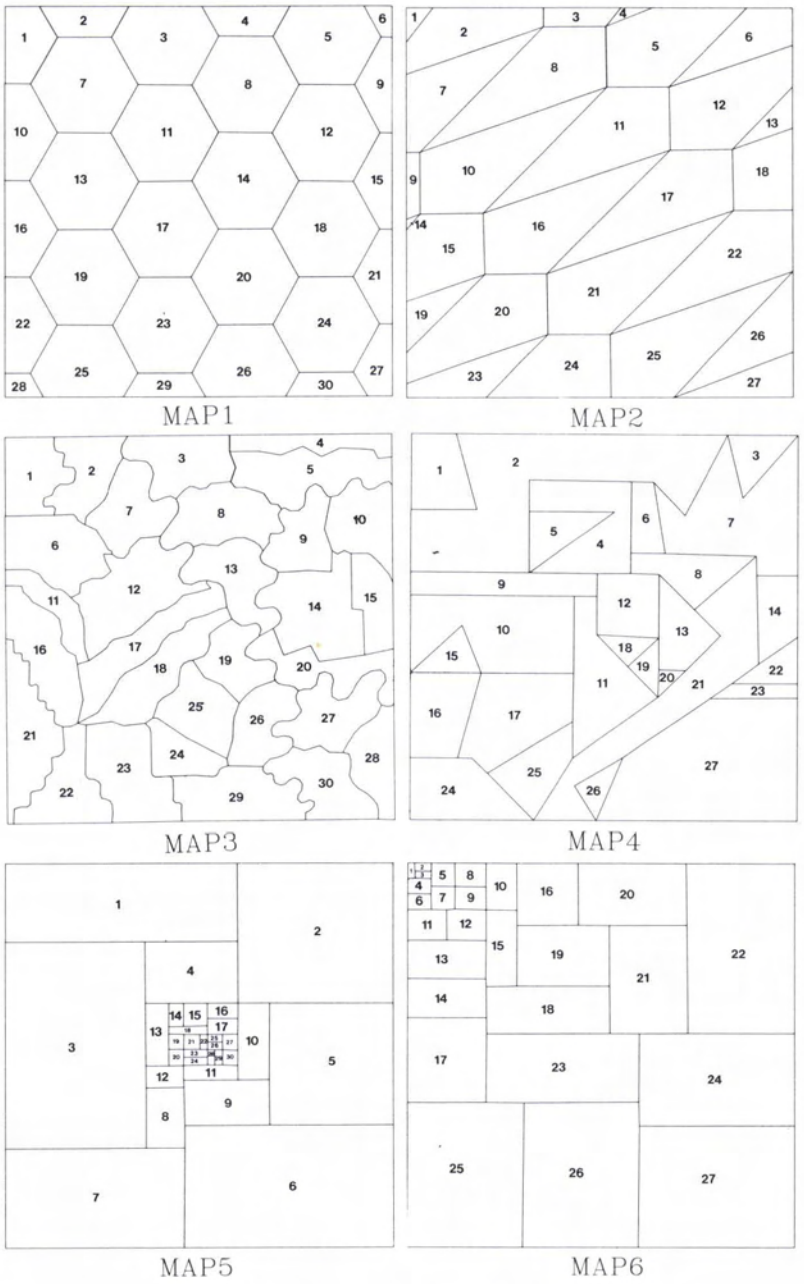

FIG. 2. Polygonal units used in the test.

tion in the sizes and shapes of the polygons. The sets of polygons were automatically superimposed onto the 50 by 50 grids. The six sets of polygons and the five test surfaces gave a test data set of 30 polygon-based maps. These maps were then used as input to the five selected polygon to grid conversion techniques. The overlay matrix was also computed as a control. The test consisted of generating 30 gridded maps, involving 855 polygons and 75,000 pixels for each of the six techniques.

Because the "target" surface of the conversion without the intermediate step of sampling by polygon was known, and because the error due to grid sampling was the same for all techniques, the conversion error could be computed by subtraction (Figure 3). Thus, for each technique, error was tab-

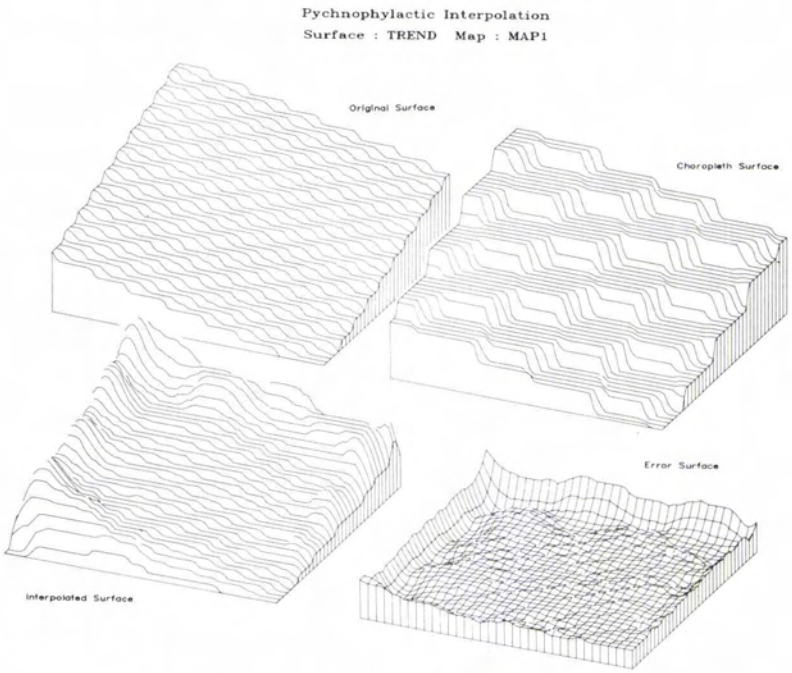

FIG. 3. Method of computing conversion error.

ulated once per map, once per polygon, and once per cell. In addition, two error measures were used. Absolute error was defined as the observed grid value after conversion minus the expected or true value. Relative error was stated in terms of percentages of expected variance for the whole map and for each polygon, and for straight percentage of true value for the cell level.

RESULTS

Table 1 shows absolute (e_a) and relative (e_r) error, and the standard deviations of error for the appropriate number of observations, structured by interpolation model. While average absolute error for all techniques was similar, relative error peaked at the polygon level. At this level, variance within polygons was over-estimated by a factor of two, while at the map and cell levels the variance was under-estimated by the same amount. The technique with the lowest absolute error and the lowest standard deviation of error was pycnophylactic interpolation. In terms of relative error, again pycnophylactic interpolation estimated surface variance most accurately. From Table 1, the techniques were ranked by performance for the 12 error statistics. These statistics were the amount, standard deviation, and range of error at the map, polygon, and cell levels. Ideal rankings were lowest absolute error, relative error closest to 100 percent of original variance, and the lowest standard deviations. Pycnophylactic interpolation ranked first, scoring 21, while the control, inverse-fourth, smoothing, inverse-squared, and trend-projection methods followed with scores of 38, 39, 40, 49, and 60, respectively.

While aggregate performance is a useful means of comparing polygon to raster conversions, some specific performances were significant. The range of performances behind the aggregate results was high. At the map level, absolute error ranged from 0.06 units, in the case of Map 6 with the linear trend using trend projection, to 39.65 units in the case of Map 5 and the exponential peak using the same method. In both of these cases shape remained constant (rectangles) while area varied as a function of surface value. Clearly, when a linear trend is an ideal model of the true surface, the trend projection model interpolates very well. When the trend was non-linear, the interpolation was the worst by a wide margin, though the inverse-squared method performed poorly on the same surface. The only technique with a low error level for this extreme combination of surface and data units was pycnophylactic interpolation.

For relative error at the map level, again the range was large. For the inverse-squared distance weighting method's estimate of the randomized harmonic surface on Map 5, only 7.7 percent of the total surface variance was captured. At the other extreme, for the exponential peak on the same map using the trend projection method, variance was estimated as 214.34 percent of the original.

TABLE 1. ABSOLUTE AND RELATIVE ERROR FOR TEST POLYGON TO GRID CONVERSIONS

Level	Invrse Sqred.	Invrse Fourth	Trend Proj.	Smooth	Pycno.	Cntrol.	Mean
			Absolute Error				
MAP							
e_a	+5.606	4.901	6.082	4.602	4.646	4.911	5.125
POLY							
e_a	5.124	4.739	5.365	5.978	4.768	4.981	5.159
CELL							
e_a	5.606	4.902	6.082	4.603	4.646	4.913	5.125
MAP							
St.Dev.	6.13	4.01	7.98	3.36	3.02	3.37	
POLY							
St.Dev.	10.83	10.40	11.02	20.39	9.03	11.54	
CELL							
St. Dev.	15.99	14.98	16.24	14.79	13.32	14.87	
			Relative Error				
MAP							
e_r	+44.66	51.36	63.99	43.19	67.18	53.19	53.93
POLY							
e_r	221.17	152.89	381.55	187.64	145.58	96.71	197.56
CELL							
e_r	125.98	55.99	88.37	4.41	91.10	4.00	61.64
MAP							
St.Dev.	33.86	34.63	47.02	29.01	46.13	31.04	
POLY							
St.Dev.	920.0	393.1	2314.6	479.5	149.0	59.0	
CELL							
St.Dev.	1689.7	1574.8	1732.2	1549.2	1408.0	1566.1	

At the polygon level, the largest absolute polygon error was achieved by the smoothing technique's estimate for polygon number 24 (Figure 2), again on Map 5, for the exponentially peaked surface. This polygon covered the peak itself, and was smoothed considerably, drastically underestimating the actual value. On the other hand, many polygons were estimated with zero error, both in absolute and relative measures. The inverse-fourth interpolation had the lowest mean polygon level absolute error, while the smoothing technique had the highest.

At the cell level, the smoothing and pycnophylactic techniques both had similarly low mean absolute error levels, while again the trend projection method had the highest. The smoothing method and the control, however, had very high relative (percent of actual) error, the best performance being that of pycnophylactic interpolation. Again, in some cases, individual cell values were estimated with zero error, but, as might be expected, the range of error was high at the cell level. In one case, the estimated value was a 517 fold under-estimate of the actual value.

In addition to the tabulation of error statistics, an attempt was made to derive an error-predictor model. Error was hypothesized to be a function of (1) the size and shape of the polygonal data collection units, as proposed by Hsu and Robinson (1970); (2) the conversion model, as shown by Morrison (1971); and (3) the complexity of the true underlying surface. The model was tested at the polygon level, where error and error variance seemed to be concentrated. Fourteen variables relating to polygon size and shape and surface complexity were used to predict absolute error in a multiple regression. The resultant coefficients of determination, stratified by model, ranged from 0.797 to 0.979, with the highest level of prediction in the direct overlay case. This implies that 97.9 percent of the variation in absolute polygon to grid conversion error is predictable, given the characteristics of the polygons and the underlying surface. When five of the variables reflecting surface complexity were excluded from the regressions, the coefficients of determination fell to the 0.161 to 0.433 range. This reflects a case where no information is available about the true surface, the most common actual situation. Again, the best performance of the error model was in the case of the direct overlay method, where 43.3 percent of the variation in absolute error was predicted by

characteristics of the polygons such as area, shape, mean perimeter length, and the nearest neighbor statistic for the polygon centroids.

Implications

Many implications for geographic data processing arose from this study. First, significant differences in polygon to grid conversion error do exist between interpolation models, and between different types of surface. In general, the best models are those which operate on the direct overlay matrix, i.e., the areal interpolation models. Direct overlay itself, which was included as a control with the expectation that all models would do better, frequently outperformed the other methods. The specific results imply that, before converting continuous spatial variables to a grid, the investigator should ask whether or not some particular type of surface was anticipated. Actual knowledge of the autocorrelation function is the ideal, but *a priori* expectations on the basis of theory or even intelligent guesswork are adequate. Clearly, the best technique to use is the one whose underlying model best matches the actual form of the "true" surface. When an expected result is matched to a model (the linear trend to the trend projection model, for example), the error is extremely low. Unfortunately, when the surface and the model are poorly matched, the error can be alarmingly high, certainly enough to invalidate GIS data layers and the analyses and decisions dependent upon them.

When nothing is known about the underlying surface, pycnophylactic interpolation or direct overlay should be used, because these models perform best on the average over the full range of surfaces, or at least the five used in this research. This finding should not be taken as an endorsement for direct overlay, which is the most frequently used method, because in most specific instances other methods can produce superior results.

In terms of the factors contributing to error, the characteristics of the polygons play a secondary role in polygon to grid conversion error. Their effect is to reinforce in their extremes the good and bad performance aspects of the conversion models. While particular combinations of polygon sizes, shapes, elongations, orientations, and perimeter lengths can contribute significantly to error, this is usually in the cases where the extremes of the polygon's characteristics coincide with the extremes of the underlying surfaces. While models of error based on polygon characteristics are able to predict only a small amount of error, they do provide a means of estimating its expected location and quantity.

For GISs, the implications are two-fold. First, the currently popular direct overlay model, while crude, frequently seems to produce the best results. Because many GISs work on the principle of direct overlay, and large amounts of data have been converted using the model, this is encouraging. However, investigators should be aware that converted data necessarily include error, as will all further transformations of the resultant GIS data layer. Second, a GIS should allow the user the flexibility to choose between conversion techniques where additional information about the true surface is known or suspected. At the very least, a user should be aware that options exist for the reduction of gridding error, and can be used to improve the quality of converted GIS data layers.

Finally, the conclusions of this study should be seen in the context of the limitations of the study. Comparatively few models, surfaces, and polygons were used, and only one grid size. The surfaces and polygons used may not be a good subset of those used for typical GIS work. Similarly, many of the findings may be specific to the data sets assembled for this study. Clearly, additional work is necessary to fully understand polygon to grid conversion error, preferably using real world rather than theoretical data. The data set assembled for the study was used to test comparatively few of the possible hypotheses on error distribution. The data for the map and polygon levels have been published (Clarke, 1982), and the author would welcome their use to elaborate upon or disprove the findings of this research.

References

Clarke, K. C., 1982. *Geographic Enhancement of Choropleth Data*, Ph.D. Dissertation, University of Michigan, University Microfilms.

———, 1984. Two-Dimensional Fourier Interpolation for Uniform Area Data, *Technical Papers, 50th Annual Meeting ASP*, 2, pp. 835-845.

Cox, N. J., and D. W. Rhind, 1978. Networks in Geographical Information Systems: A British View, *Proceedings, First International Advanced Study Symposium on Topological Data Structures for Geographic Information Systems*, Cambridge, Mass., Harvard Lab. for Computer Graphics and Spatial Analysis.

Crain, I. K., 1970. Computer Interpolation and Contouring of Two-Dimensional Data: A Review, *Geoexploration*, 8, 71-86.

Estes, J. E., 1982. Remote Sensing and Geographic Information Systems Coming of Age in the Eighties, in B. F. Richason, (ed.) *Proceedings, Pecora VII Symposium*, Falls Church, VA., ASP.

Goodchild, M. F., and N. S. Lam, 1980. Areal Interpolation: A Variant of the Traditional Spatial Problem, *Geo-Processing*, 1, 297-312.

Harbaugh, J. W., and D. F. Merriam, 1969. *Computer Applications in Stratigraphic Analysis*, J. Wiley, New York.

Hsu, M. L., 1975. Filtering Process in Surface Generalization and Isopleth Mapping, in J. C. Davis, and M. J. McCullagh, *Display and Analysis of Spatial Data*, J. Wiley, London, pp. 115–129.

Hsu, M. L., and A. H. Robinson, 1970. *The Fidelity of Isopleth Maps*, Minneapolis, University of Minnesota Press.

Jenks, G. F., 1963. Generalization in statistical mapping, *Annals, Association of American Geographers*, 53, 15-26.

Morrison, J. L., 1971. *Method-Produced Error in Isarithmic Mapping*, Washington D.C., ACSM Cartographic Division.

Muller, J., 1977. Map Gridding and Cartographic Errors/ A Recurrent Argument, *The Canadian Cartographer*, 14, 2, 152-167.

Nordbeck, S., and B. Rystedt, 1970. Isarithmic Maps and the Continuity of the Reference Interval Functions, *Geografiska Annaler*, B2, 92-123.

Peucker, T. K., et al., 1978. The Triangulated Irregular Network, *Proceedings, ASP Digital Terrain Models Symposium*, St. Louis, Mo., pp. 516–540.

Peuquet, D. J., 1981. An Examination of Techniques for Reformatting Digital Cartographic Data/ Part 2: The Vector to Raster Process, *Cartographica*, 18, 3, 21-33.

Rhind, D. W., 1971. Automated Contouring—An Empirical Evaluation of some Differing Techniques, *The Cartographic Journal*, 8, 2, 145-158.

Robinson, A. H., J. B. Lindberg, and L. W. Brinkman, 1961. A Correlation and Regression Analysis Applied to Rural Farm Population Densities in the Great Plains, *Annals, Association of American Geographers*, 51, 211-221.

Sampson, R. J., 1978. *The Surface II Graphics System*, Lawrence, Kansas, Kansas Geological Survey, Revised.

Shepard, D., 1968. A Two-Dimensional Interpolation Function for Irregularly Spaced Data, *Proceedings, Association for Computing Machinery*, pp. 317-323.

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

Walters, R. F., 1969. Contouring by Machine: A Users Guide, *Bulletin, American Association of Petroleum Geologists*, 35, 11, 2324-2340.

Zobrist, A., 1977. Elements of an Image Based Information System, *IEEE Workshop on Image Database Management*, pp. 55-63.

(Received 24 March 1984; revised and accepted 25 January 1985)
