CHOROPLETH MAP ACCURACY: CHARACTERISTICS OF THE DATA

Alan M. MacEachren
Department of Geography
Virginia Polytechnic Institute
and State University
Blacksburg, VA 24061

ABSTRACT

The accuracy of maps based on aggregate data is dependent upon (a) the extent to which aggregate values calculated for each enumeration unit are representative of the entire unit and, (b) the data classification system used to assign these values to classes. Size and compactness of these units as well as the variability of the distribution mapped are important factors in determining the accuracy of aggregate values calculated for the units. In this study, the individual and relative importance of these enumeration unit and surface characteristics are examined. Analysis indicates that all three variables exert a significant influence on accuracy of aggregate values with surface variation accounting for the greatest portion of the variation in accuracy.

INTRODUCTION

In contrast to the considerable attention directed toward the accuracy with which choropleth maps communicate spatial information, little attention has been given to the accuracy of the maps themselves. With the advent of geographic information systems and the concomitant development of interactive choropleth mapping capabilities such as the Domestic Information Display System and the SASGRAPH package, choropleth maps are becoming increasingly important in planning and decision making. For maps to be effectively and appropriately used in this context, it is necessary that accuracy of these maps be carefully considered.

Choropleth map accuracy is, to a large degree, a function of methods by which data are organized. Data for choropleth maps consist of aggregate values for enumeration units such as states, counties, or census tracts. An underlying assumption of the choropleth technique is that data within each unit are of equal value and evenly distributed across the unit. For choropleth maps, then, accuracy will be a function of (a) the variation of data within each unit from the aggregate value representing that unit and, (b) the data classification system used to assign values to classes.

While the effect of data classification procedures on the accuracy with which aggregate values are represented has been considered (Jenks and Caspall, 1971 and Monmonier, 1982), the influence of data characteristics and organization procedures on aggregate value accuracy has been largely ignored. One explanation for a lack of attention to accuracy of the data may be a perceived inability to manipulate the variables involved. While cartographers can control shading patterns or data classification procedures on a choropleth map, they have little or no ability to control size and shape of enumeration units or the nature of the distribution mapped.

Whether or not the cartographer can control all variables, a responsibility exists to evaluate the potential for error on maps produced. In some cases this evaluation may result in a decision that the available aggregate data will not produce a sufficiently accurate map or that an alternative to a choropleth map should be used. In other cases, when it is decided that the map is to be constructed, map users could be provided with a measure of overall map accuracy or alerted to regions of the map where, due to questionable data, caution should be taken in interpretation.

As a step toward development of a method for determining the potential for error of individual choropleth maps, the focus of the present study is on the correspondence between aggregate values and the data they represent. It is postulated that three factors: enumeration unit size, enumeration unit compactness, and variability of the data distribution, are the determinants of aggregate value accuracy. Variability of the data distribution is expected to be indicative of data variation within each unit. Therefore, accuracy will decrease as distribution variability increases.

Size and compactness of units are expected to influence aggregate value accuracy because of their direct correspondence to distances among individual data elements. The larger and less compact the units, the farther apart individual locations within the units will be and, consequently, the more likely it is that their characteristics will vary. Aggregate values representing units, therefore, will decrease in accuracy as size of units increases and as compactness decreases.

Coulson has suggested that size and compactness of units have an equal influence on accuracy of aggregate values. From a theoretical point of view, this is readily apparent. In practice, however, the relative importance of size and compactness will be a function of the variation of each factor for the units involved.

METHODOLOGY

The focus of the present study was on one aspect of choropleth map accuracy -- the accuracy of aggregate values to be mapped. For this purpose, a set of contiguous enumeration units, such as the counties in a state, was not essential. In an effort to obtain an adequate range in size and shape of units, individual rather than contiguous enumeration units were used. The units selected consisted of a stratified random sample of six counties from each of nine regions of the U.S. (Fig. 1a and 1b). The actual units varied in size by a ratio of about 6 to 1 from largest to smallest. For convenience of illustration, however, all units are scaled to the same area.

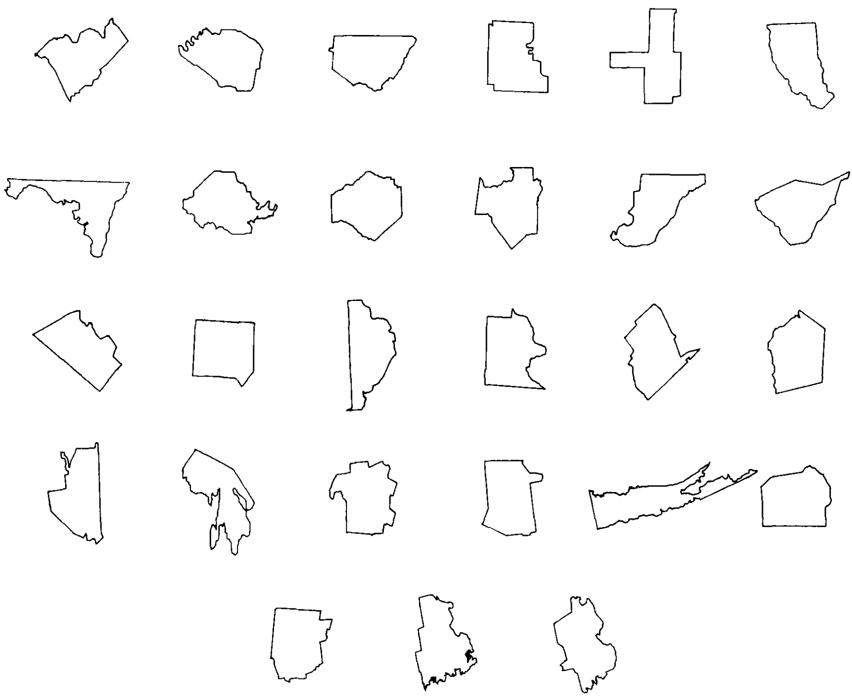

Fig. 1a. Sample Units

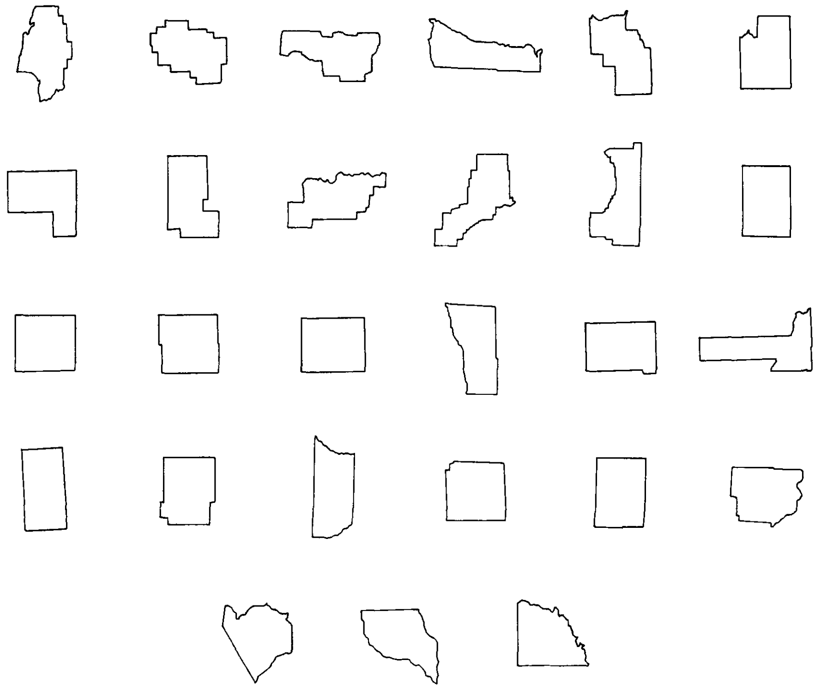

Fig. 1b. Sample Units

For the influence of data distribution variability on aggregate value accuracy to be examined, distributions representing a range in variability were necessary. Four distributions were utilized. The first (Fig. 2a) was a simple linear surface that decreases in value along a diagonal. This was assumed to be the simplest surface. The remaining distributions were derived from actual topographic surfaces and can be described as: a roughly conic surface (Fig. 2b), an undulating linear surface (Fig. 2c), and a highly irregular surface (Fig. 2d). Each surface was generated from a set of control point values by the Surface II Graphics System (Sampson, 1975). This system generates a square grid matrix of z-values from which an isoline map or perspective plot can be created. In this case each matrix consisted of 112 rows and 75 columns 1/10 of an inch apart.

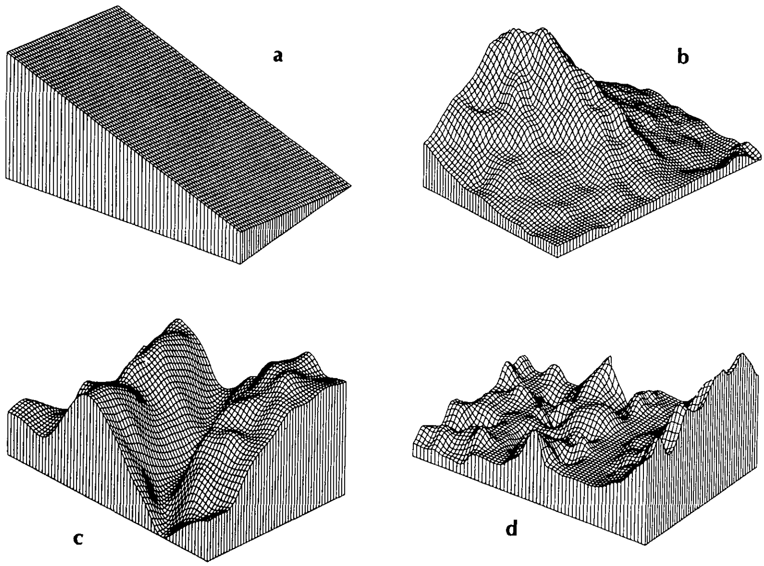

Fig. 2 Sample Surfaces

Measurement of Accuracy

In terms of the choropleth assumption of homogeneity within units, accuracy can be measured as the variance of values occurring within a unit around the mean or aggregate value used to represent that unit. To obtain this measure, each unit was positioned, at a random location and orientation, on the grid matrix representing a distribution. Points of the matrix inside the unit were determined (Fig. 3). There were between 30 and 300 points within each unit depending on its size. The mean and standard deviation of z-values at these points were calculated and the coefficient of variation for the standard deviation was computed. This coefficient of variation was used as the measure of aggregate value accuracy.

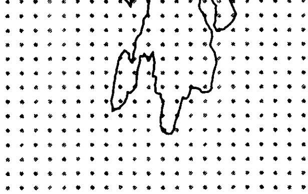

Fig. 3 Sampling of Grid Matrix

Measurement of Variables

A variety of methods have been proposed for the measurement of enumeration unit compactness (Blair and Biss, 1967). While a number of simple measures based on the perimeter or area of the unit exist, the method that, from a theoretical standpoint, should be most accurate deals with the unit as a whole rather than with a single parameter of the unit. Each unit is considered to be composed of a series of infinitesimally small elements of area (Fig. 4). Variation in location of these elements in relation to the unit's centroid is the basis for the measure. It is calculated as the sum of the variance in X and Y locations of the elements, adjusted so that values range from zero to one, the latter being the value for a circle, the most compact shape. Versions of this measure have been presented by Bachi (1973), Blair and Biss (1967), and Coulson (1978). The relative distance standard deviation is the form used here.

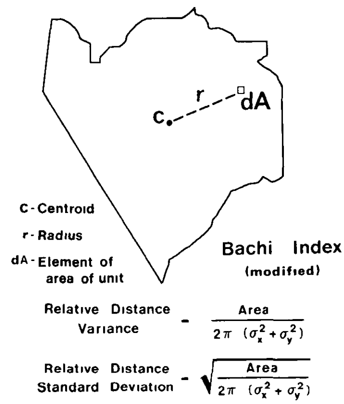

Fig. 4 Compactness Index

While compactness can be compared to the most compact possible shape, there is no single standard to which size can be compared. Any measure of size devised, therefore, is a relative measure; meaningful only in a given context. One practical measure of size is to calculate a ratio between the size of each unit and an arbitrary standard. Convenient standards are the largest, smallest, mean, or median size of units. The largest and smallest units have the advantage of resulting in a scale from zero to one.

Coulson (1978) has advocated the use of the smallest unit as a standard in order to produce a scale comparable to that for compactness. Values would range from zero to one with high values indicating a potentially more accurate aggregate value, as they do for the standard distance deviation scale of compactness. The problem with this approach is that the size ratio is dependent on one, possibly extreme, value.

Although not as easily interpreted, the median size of the set of units will provide a more stable standard and is used here. The median is preferable to the mean because the distribution of enumeration unit size is likely to be highly skewed with a small number of large units that would exert unwarrented influence on the mean.

Surface variability can be measured in a number of ways. An important consideration in selecting a measure is the frequency of variation. For example, a distribution that exhibits extreme variability on a continental scale may exhibit little or no variation across a possible mapping unit (e.g., a county). A measure of data distribution variation that takes mapping unit size into account is requisite.

In the present study, surface variability is measured by comparing neighboring z-values in the grid matrix representing each distribution. The measure used is the spatial autocorrelation of grid z-values at a lag equal to the average longest axis of the units examined. This measure will reflect the maximum likely variation within an average mapping unit.

ANALYSIS

Previous research (MacEachren, 1982) has demonstrated that both size and compactness of enumeration units exhibit the expected influence on aggregate value accuracy. Accuracy increases as size decreases and as compactness increases. The influence of each factor on aggregate value accuracy was shown to be a function of the factor's variation across the units examined; the greater the variation the greater the influence on accuracy.

In the present study, a third factor, data distribution variability is considered. Multiple regression analysis is used to determine the relative influence of unit size, unit compactness, and distribution variability on aggregate value accuracy.

To examine the influence on accuracy of these factors, the accuracy value, the coefficient of variation, is calculated for each of the 54 units positioned on each of the four distributions. Multiple regression of these variation ratios with the measures of unit size, unit compactness, and data distribution variability indicates that all three factors explain a significant portion of the variation in aggregate value accuracy (Table 1).

TABLE 1. MULTIPLE REGRESSION OF ACCURACY WITH SIZE, COMPACTNESS, AND SURFACE VARIATION

Variables	Multiple R	R^2	R^2change	Simple R
Surface variation	.76	.58	.58	.76
Size	.94	.89	.32	.56
Compactness	.97	.93	.04	-.20

Data distribution variability, as measured by spatial autocorrelation, provides the greatest contribution to an explanation of variation in aggregate value accuracy. As expected, with increasing variation at a frequency corresponding to enumeration unit size, there is a decrease in aggregate value accuracy.

For the enumeration units included, size exhibits greater variation than does compactness (Table 2). As expected, therefore, unit size provides the greater contribution toward explaining variation in accuracy. This is evident in both the simple correlation of the variables with accuracy and in their respective contributions to the multiple regression.

TABLE 2. SIZE AND COMPACTNESS COMPARISON

	Mean	S.D.
Square Root of Size Ratio	0.96	0.28
Compactness Index	0.88	0.09

CONCLUSIONS

The specific focus of the present study has been the relative influence on aggregate value accuracy of enumeration unit size, enumeration unit compactness, and data distribution variability. Results indicate that data characteristics have a greater influence on aggregate value accuracy than do characteristics of the enumeration units to which data are assigned. Enumeration unit characteristics, however, have also been shown to be significant factors.

These findings suggest that, while the extent to which data meet choropleth assumptions remains a primary consideration for choosing the choropleth technique, unit size and compactness should be considered as well. It is possible, for example, that while a particular phenomenon is well suited to choropleth representation, the size and compactness of the units to which data are aggregated may produce significant differences in accuracy from one part of the map to another.

Results of the present study are one step toward the overall goal of a method for determining potential error in specific choropleth maps. The importance of both data characteristics and the manner in which data are aggregated have been demonstrated. To produce maps of potential error in specific choropleth maps, however, the relative importance of these variables and data classification procedures must be determined. In addition, a method of estimating data distribution variability from aggregate data when individual data are not available must be derived.

Developments in both hardware and software of computer-assisted cartography are resulting in an increased potential for the use of maps in decision making. It is now possible to produce maps of current information quickly and inexpensively. As thematic maps are increasingly used to make decisions rather than simply illustrate decisions, more careful consideration of their accuracy is essential.

REFERENCES

Bachi, R. 1973, "Geostatistical Analysis of Territories," Bulletin: International Statistical Institute, (Proceedings of the 39th Session), Vol. 45, Book 1, 121-131.

Blair, D. J. and T. H. Biss 1967, "Measurement of Shape In Geography: An Appraisal of Methods and Techniques," Bulletin of Quantitative Data for Geographers, No. 11.

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, Vol. 60B, 48-64.

Jenks, G. F. and F. C. Caspall 1971, "Error on Choropleth Maps: Definition, Measurement, Reduction," Annals of the Association of American Geographers, Vol. 61, 217-244.

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, 512-521.

Monmonier, M. S. 1982, "Flat Laxity, Optimization, and Rounding in the Selection of Class Intervals," Cartographica, Vol. 19, 16-27.

Sampson, R. J. 1975, SURFACE II Graphics System, Series on Spatial Analysis, No. 1, Kansas Geological Survey.