ANNALS of the
Association of American Geographers

Volume 61 June, 1971 Number 2

ERROR ON CHOROPLETHIC MAPS: DEFINITION, MEASUREMENT, REDUCTION

GEORGE F. JENKS AND FRED C. CASPALL

ABSTRACT. Communication, whether oral, written, or graphic depends upon the ability of one individual to transfer information to another. In this essay we have tried to illustrate the fact that error in choroplethic mapping inhibits the transfer of information and that there are methods for improving this type of map as a communicative tool. We have done this by first defining overview, tabular, and boundary map uses. Second, techniques for the measurement of the error components of these three uses have been developed. Third, new reiterative and forcing manipulative techniques for choroplethic map data processing have been evolved. Lastly, the relationship between map accuracy and the information carrying capacity of a choroplethic map is set forth in hypothetical terms.

MAP-MAKERS and map-readers have long expressed concern about the inaccuracies inherent in choroplethic maps.[1] A choroplethic map is a generalization of the reality of an areal distribution and as such it must contain error. Writers agree that some generalizations are more accurate than others, and that the cartographer should present his reader with the best representation that can be constructed, but there are great inconsistencies in the evaluation of, and methods devised for obtaining accurate, choroplethic maps. Schultz suggested that the map-maker should strive for visual attractiveness and easily understood areal patterns.[2] Jones suggested that class breaks should be selected at "critical" values which may be derived from field observations or, as is often the case, from a particular known or unknown bias held by the map-maker.[3] The majority of writers stress the need for objectivity in the generalizing process, and to achieve this aim they suggest various methods for manipulating the data into "accurate" classes.[4] Recently Armstrong called for

Accepted for publication 17 March 1970.

Dr. Jenks is Professor of Geography at the University of Kansas in Lawrence and Dr. Caspall is Assistant Professor of Geography at Western Illinois University in Macomb.

[1] The derivation of the term choropleth and an analysis of some of the problems in using this form of symbolization are given in J. K. Wright, "The Terminology of Certain Map Symbols," *The Geographical Review,* Vol. 34 (1944), pp. 653–54, and J. K. Wright, "Map Makers are Human," *The Geographical Review,* Vol. 32 (1942), pp. 527–44.

[2] G. M. Schultz, "An Experiment in Selecting Value Scales for Statistical Distribution Maps," *Surveying and Mapping.* Vol. 21 (1961), pp. 224–30.

[3] W. D. Jones, "Ratios and Isopleth Maps in Regional Investigation of Agricultural Land Occupance," *Annals,* Association of American Geographers, Vol. 20 (1930), pp. 177–95, especially page 181.

[4] Discussions of this type may be found in J. W. Alexander and G. A. Zahorchak, "Population-Density Maps of the United States: Techniques and Patterns," *Geographical Review,* Vol. 33 (1943), pp. 457–66; J. R. Mackay, "An Analysis of Isopleth and Chloropleth Class Intervals," *Economic Geography,* Vol. 31 (1955), pp. 71–81; and M. W. Scripter, "Nested-Means Map Classes for Statistical Maps," *Annals,* Association of American Geographers, Vol. 60 (1970), pp. 385–93.

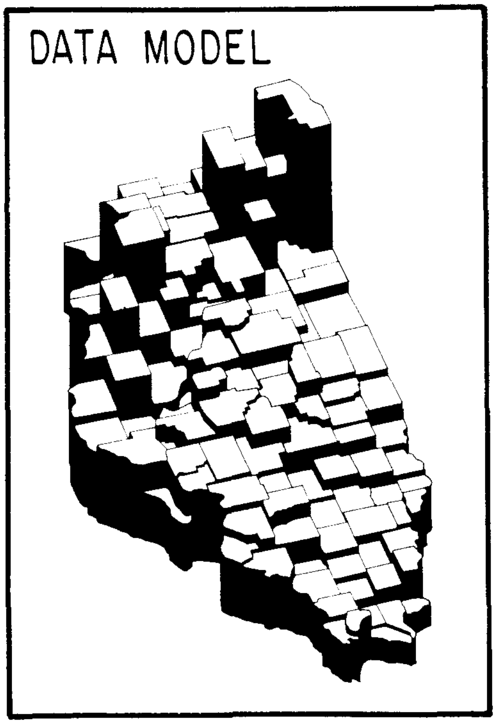


FIG. 1. A three-dimensional map of a statistical distribution. This data model is created by visualizing each enumeration unit as the base of a prism whose height is proportional to the intensity value of the mapped phenomenon.

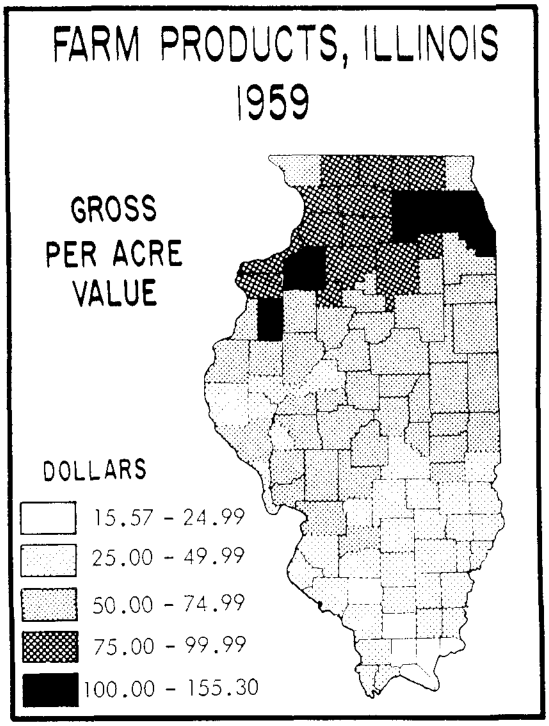


FIG. 2. A choroplethic map is a planimetric generalization of a geographical distribution. This map is a generalization of the data model in Figure 1.

standardized classes based upon the statistical distribution of the data about the mean.[5]

It appears to us that one very serious deficiency is common to all of these discussions, for rarely is the map-reader and the way he obtains information from a map considered. Maps are communicative devices which perform a myriad of functions, and the serious map-reader commonly uses a map for one or more of three purposes. First, he may seek an overview of the statistical distribution from a choroplethic map, much as he obtains the "lay of the land" from a topographic map.[6] Second, he may think of the statistical map as an areal table which he uses as a source of specific information about a place. Lastly, he may focus upon the boundary lines between patterns or shadings and compare these boundaries with those that are held as mental images or those which occur on other maps. If map-readers do, in fact, use choroplethic maps for these purposes (overview, tabular, and boundary), it is clear that the utility of such a map should be measured in light of these needs.

In this paper we investigate choroplethic map construction in light of map-user requirements. To achieve this objective we first inquire into the nature of statistical distributions. Second, several traditional generalizing techniques are reviewed. Third, the character of error in choroplethic mapping is defined, and suitable methods for measuring error are presented. Finally, new methods of creating choroplethic maps are suggested.

THE VOLUMETRIC NATURE OF INTANGIBLE DISTRIBUTIONS

Abstract statistical phenomena are unlike tangible features of the physical environment

[5] R. W. Armstrong, "Standardized Class Intervals and Rate Computation in Statistical Maps of Mortality," *Annals*, Association of American Geographers, Vol. 59 (1969), pp. 382–90.

[6] Statistical distributions, as the term is commonly used in geography, are distributions which are derived and mapped from enumerated data. Census data are enumerated data and maps made from these "statistics" represent statistical distributions.

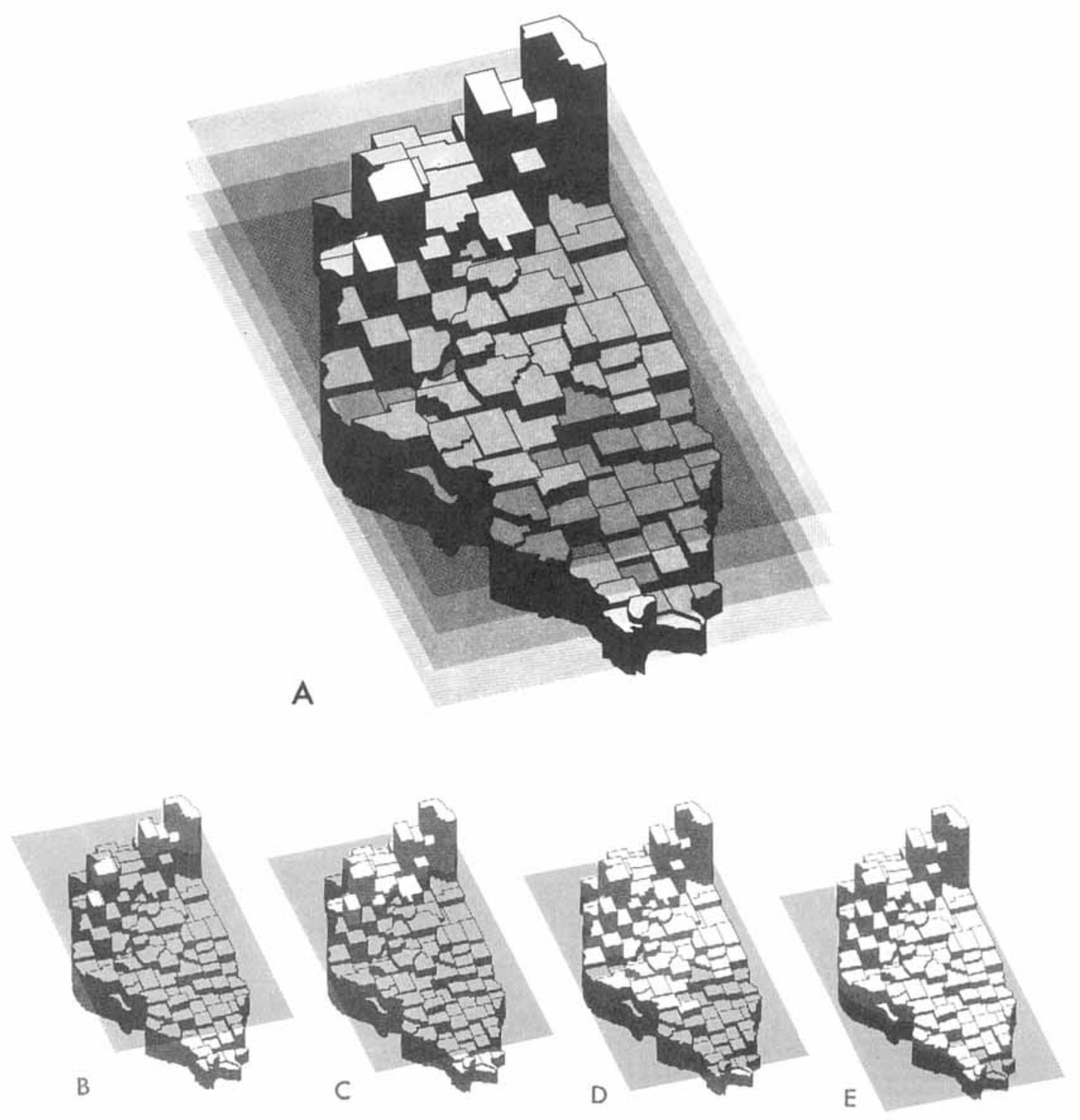


FIG. 3. The classes for a choroplethic map are derived by cutting a data model into layers with a series of parallel planes. (A) Four planes are utilized to create a five class map; in this case they are elevated at $100 (B), $75 (C), $50 (D), and $25 (E). The cut lines on these data models (B, C, D, and E) provide the boundaries for the classes on the final planimetric representation.

in that they cannot be seen. This means that the map-maker and the map-reader perceive them by a rational rather than a sensory process. Furthermore, neither can rely on "real world" experience as a guide to understanding. We can, however, visualize each enumeration unit area as the base of a prism whose height is proportional to the intensity value of the mapped phenomenon. When a series of such prisms are constructed tangentially, they form a stepped surface (Fig. 1).[7]

Choroplethic maps are planimetric representations of volumetric statistical distributions, and they are normally symbolized by patterns which divide the area into subregions (Fig. 2).

[7] For a more detailed discussion of stepped surface volumes see C. F. Schmid and E. H. MacConnell, "Basic Problems, Techniques, and Theory of Isopleth

Each choroplethic map class represents a layer or vertical segment of the statistical volume. These classes are obtained by cutting the model with a series of planes parallel to the datum plane (Fig. 3A). Four cutting planes are used to create a five class map (Fig. 2). The intensity value (elevation) of each cutting plane is normally given as the first class value in the map legend. The cutting planes used to create the map in Figure 2 had intensity values of $25, $50, $75, and $100; each is shown separately in Figures 3 B, C, D, and E. All of the enumeration units which lie above the $100 cutting plane (Fig. 3B) are shown as the black region on the map in Figure 2. The cutting plane delimits the boundary of this region. Similarly, each additional region is bounded by the cut line of one of the planes.

When a data set is grouped into classes and shown on a choroplethic map the individual intensity values are obscured. Instead, the reader observes a limited number of tones (one for each class) and the limiting values for the class which are conventionally given in the legend. He does not know which enumeration unit in a class is associated with either of the class limit values nor does he know the intensity values of other members of the class. One can question the soundness of this legend convention, for it can be held that the reader would be better informed if the map-maker presented the mean value of the class in the legend rather than the class limits. In Figure 2 the map legend thus might read, from lowest to highest: $19.36, $36.31, $58.94, $84.21 and $122.50. For the present it is sufficient to recognize that each class can be described by a single intensity value—the class mean.

All members (enumeration units) of a class lose their identity in a choroplethic generalization and each takes the value of the class mean, because all are symbolized with a single tone. The generalization is thus conceived as a simplification of the original distribution and in three-dimensional format it takes a volumetric form as a generalized model (Fig. 4). This generalized model of the Illinois distribution is an analog of the data model in which one hundred and two data prisms have been reduced to five larger but simplified prisms.

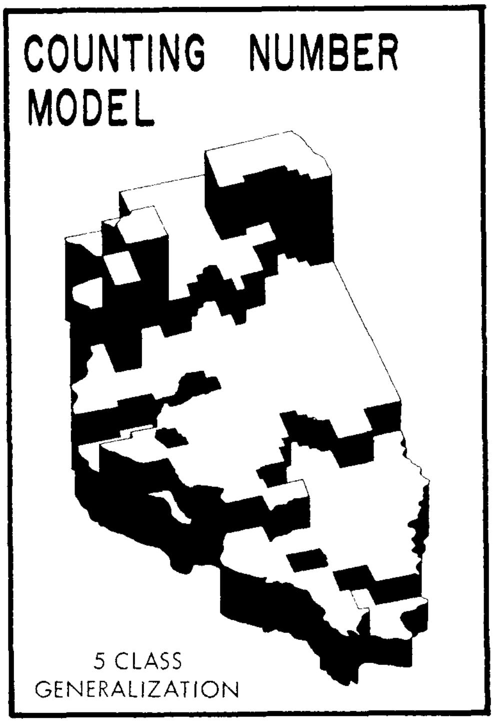


Fig. 4. This generalized model was created by constructing five prisms with heights equal to the intensity value of the mean of each class. It is a simulation of the map in Figure 2 and an analog for the data model in Figure 1.

Three dimensional models enable one to "see" an intangible distribution in a tangible or sensory format. They also provide an understanding of how choroplethic maps are created and of the form taken by the generalization. The models, however, do not solve the problems involved in classing data for choroplethic maps, since the number and location of the cutting planes are highly variable. In the following section, models display the traditional generalizing procedures to assist in evaluation of the methodology.

Traditional Construction Procedures

Robinson calls selection of the interval the most important problem a cartographer faces in employing isarithms.[8] The same thing can be said for selection of the elevations of the

Mapping," *Journal of American Statistical Association*, Vol. 50 (1955), pp. 220–39, and G. F. Jenks, "Generalization in Statistical Mapping," *Annals*, Association of American Geographers, Vol. 53 (1963), pp. 15–26.

[8] A. H. Robinson, *Elements of Cartography*, second edition (New York: John Wiley and Sons, 1960), p. 190.

TABLE 1.—PER ACRE VALUE OF GROSS FARM PRODUCTS, IN DOLLARS, BY COUNTY, ILLINOIS, 1959

Adams	47.29	Hardin	15.57	Morgan	57.26
Alexander	28.71	Henderson	68.45	Moultrie	55.44
Bond	46.13	Henry	116.40	Ogle	87.87
Boone	84.84	Iroquois	55.30	Peoria	57.49
Brown	34.63	Jackson	31.19	Perry	24.83
Bureau	85.41	Jasper	41.02	Piatt	67.04
Calhoun	34.58	Jefferson	32.14	Pike	53.77
Carroll	92.45	Jersey	52.05	Pope	17.82
Cass	50.53	Jo Daviess	50.12	Pulaski	32.22
Champaign	60.66	Johnson	18.57	Putnam	66.32
Christian	59.89	Kane	119.90	Randolph	39.84
Clark	41.20	Kankakee	62.65	Richland	31.78
Clay	32.28	Kendall	96.37	Rock Island	75.51
Clinton	50.52	Knox	72.76	St. Clair	47.21
Coles	52.21	Lake	71.61	Saline	27.16
Cook	155.30	La Salle	77.29	Sangamon	62.06
Crawford	36.24	Lawrence	33.26	Schuyler	38.22
Cumberland	50.11	Lee	86.75	Scott	45.40
De Kalb	131.50	Livingston	60.27	Shelby	49.38
De Witt	57.22	Logan	57.59	Stark	80.45
Douglas	56.35	McDonough	68.19	Stephenson	83.90
Du Page	111.80	McHenry	79.66	Tazewell	64.28
Edgar	52.89	McLean	68.25	Union	30.97
Edwards	33.83	Macon	59.65	Vermilion	57.40
Effingham	39.88	Macoupin	54.27	Wabash	38.30
Fayette	33.82	Madison	51.15	Warren	100.10
Ford	54.07	Marion	31.28	Washington	40.62
Franklin	28.20	Marshall	63.67	Wayne	32.69
Fulton	50.64	Mason	39.72	White	31.66
Gallatin	37.57	Massac	15.93	Whiteside	96.78
Greene	54.32	Menard	71.05	Will	62.41
Grundy	58.50	Mercer	75.29	Williamson	23.42
Hamilton	26.09	Monroe	40.36	Winnebago	76.43
Hancock	52.47	Montgomery	51.15	Woodford	73.52

Computed from: U.S. Bureau of the Census. Census of Agriculture: 1959. Vol. 1, Counties, Part 12, Illinois.

cutting planes for a choroplethic map. We wonder whether the average map-author realizes the significance of this task, for it is apparently performed without an understanding of its impact. As a case in point, we present as our first traditional procedure a method which may do an injustice to some of our colleagues. We are certain, however, that many maps result from an almost accidental setting of class limits.

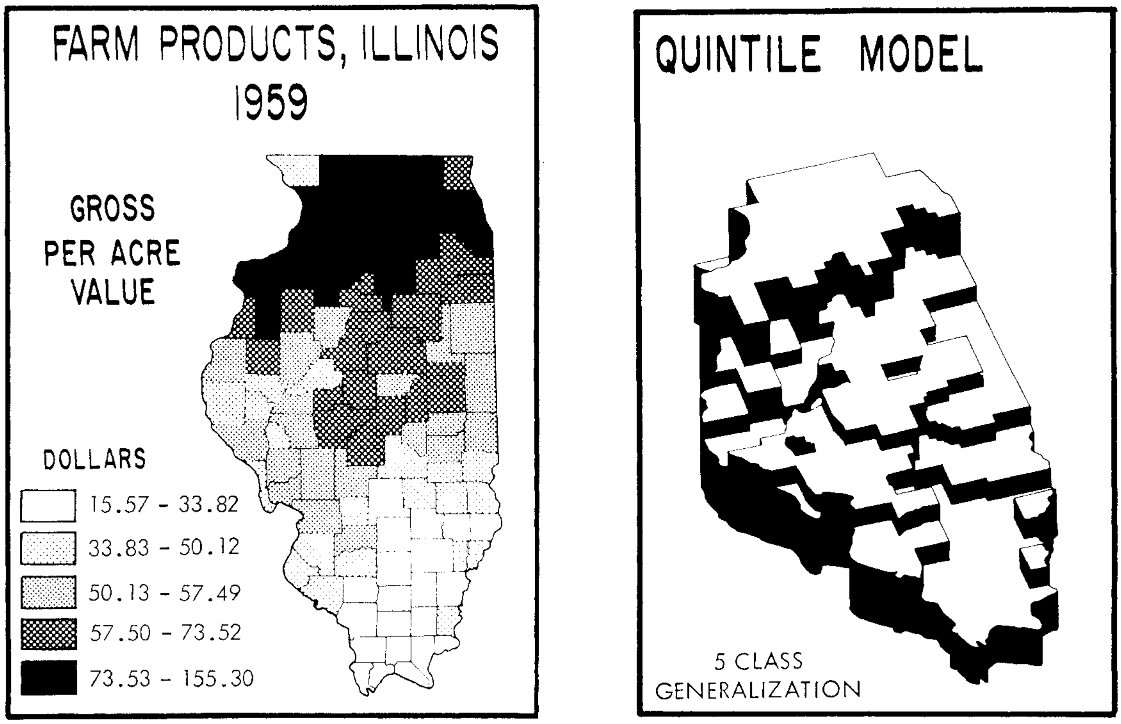


FIG. 5. This map was created by grouping the intensity values in Table 1 into subdivisions containing 21, 21, 20, 20, and 20 enumeration units each. The values are arrayed in order prior to the grouping so that the members of each class contain enumeration units with similar intensities.

Typically the unsophisticated map-author obtains a data set (such as Table 1), a base map, a set of colored pencils, and proceeds to subdivide the data into groupings. He may note, for instance, that the data range from $15.57 to $155.30, and for some reason he decides that he will create five classes. He then looks for a series of counting numbers, numbers in multiples of five or ten, which divide the data into five groups. For example, such a grouping might split the data at $25, $50, $75, and $100. When this decision is made the map is compiled, drafted, and regarded as the "truth" (Fig. 2). If data are plotted carefully the map may not be arithmetically inaccurate, but neither may it present the distribution accurately, since the map-author does not known whether the generalized model of the map approaches the data model in surface configuration, nor more importantly, is he aware of this need.

If the map-author is slightly more perceptive, he may create a number of different compilations before selecting the one he will present. He may, for example, select class limits to fit an arithmetic or geometric progression, as suggested by Wright, and by Schmid and McCannell.[9] On the other hand, he may create class breaks at "critical values" as suggested by Jones.[10] These values may be derived from field observations, or they may be more fictional than real. After these experimental maps are compiled, the map-maker selects the one which best suits his purpose. This purpose is, of course, unknown to the map-reader, but it may be the one which "presents a smoother, clearer gradation, a more easily read and remembered picture."[11] Here is an opportunity for the map-author to select a map which suits a known or unknown bias. As Schultz points out, "enough has been said, however, to show that a skilled cartographer can manipulate his map like a musician does his instrument, bringing out the quality he wants."[12]

[9] Wright (1942), op. cit., footnote 1, p. 537.

[10] Jones, op. cit., footnote 3, p. 180.

[11] Schultz, op. cit., footnote 2, p. 226.

[12] Schultz, op. cit., footnote 2, p. 230.

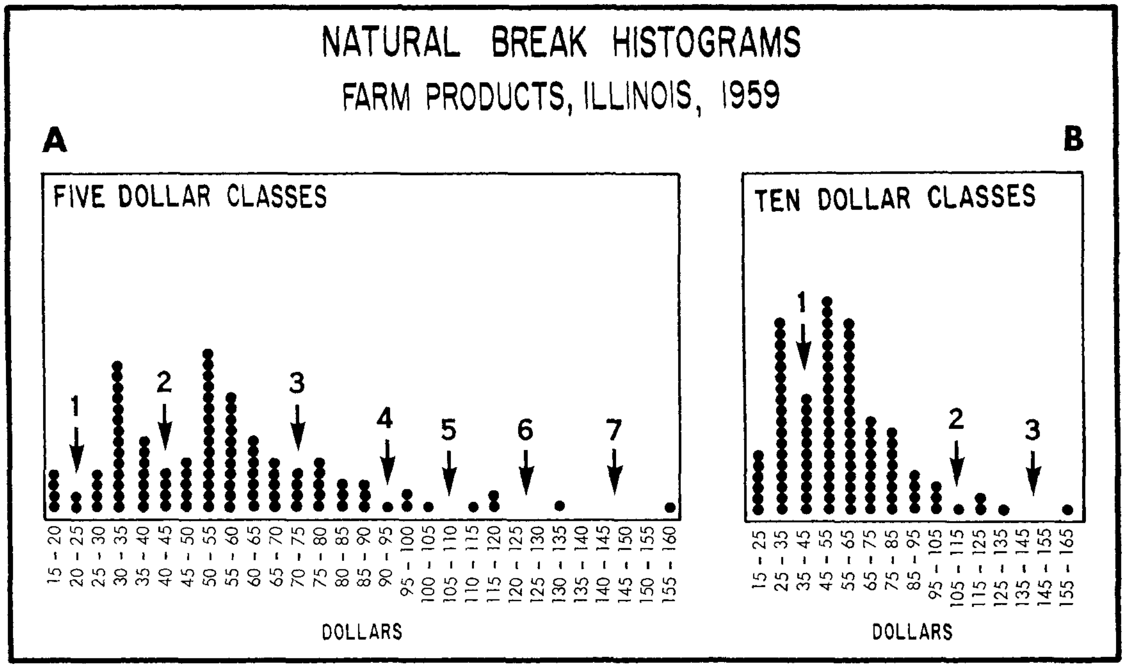


FIG. 6. Mappable data are frequently plotted on histograms so that "natural breaks" in the statistical distribution can be identified and used as class limits for a choroplethic map. These two histograms illustrate one serious problem with this technique. The number of breaks, and to a lesser degree, their location, varies with the size of the interval used to plot the data.

Other map-authors strive to avoid personal bias by manipulating their class limits into positions set forth by experienced cartographers and statisticians. They may subdivide the data into -tile groupings, following the rationale used by educational statisticians (Fig. 5), or they may follow the "natural break" theory set forth by Alexander and Zahorchak.[13] Natural break classes are obtained from histogram plots of the data, and the class limits are taken at low spots on the histogram (Fig. 6). The two histograms in Figure 6 illustrate one of the problems encountered when using this technique, because the number of breaks is related to the plotting scale on the X axis of the graph. The cartographer faces the problem of having to use four classes, eight classes, or of selecting the more important breaks to create other levels of generalization. It is clear that the histogram scale influences the position of the breaks in the data and that some subjective analysis of the plot must be made. The map in Figure 7 was created by an arbitrary selection of the first four breaks in Figure 6A.

Three additional classification techniques deserve mention here. The first attempts to fit data to a function which can be subdivided into equal parts of the general equation $y = f(x)$.[14] Scripter suggested that classes be created by using the mean of the data as the first break point. The second and third break points are determined by finding the mean of the lower and upper halves of the data values. In this manner maps with 2, 4, 8, and 16 or more classes can be constructed.[15] Mackay based a technique of classing data for statistical maps on the clinograph, which has long been used in land form analysis.[16] Class breaks are taken at the more significant changes of slope in the clinograph plot. The map in Figure 8 results from this method of classification.

Recently Armstrong has made a strong plea for a standard procedure to establish standard-

[13] Alexander and Zahorchak, op. cit., footnote 4, p. 459.

[14] G. F. Jenks and M. F. C. Coulson, "Class Intervals for Statistical Maps," *International Yearbook of Cartography*, Vol. 3 (1963), pp. 119–34.

[15] Scripter, op. cit., footnote 4, p. 803.

[16] Mackay, op. cit., footnote 4, pp. 73–78.

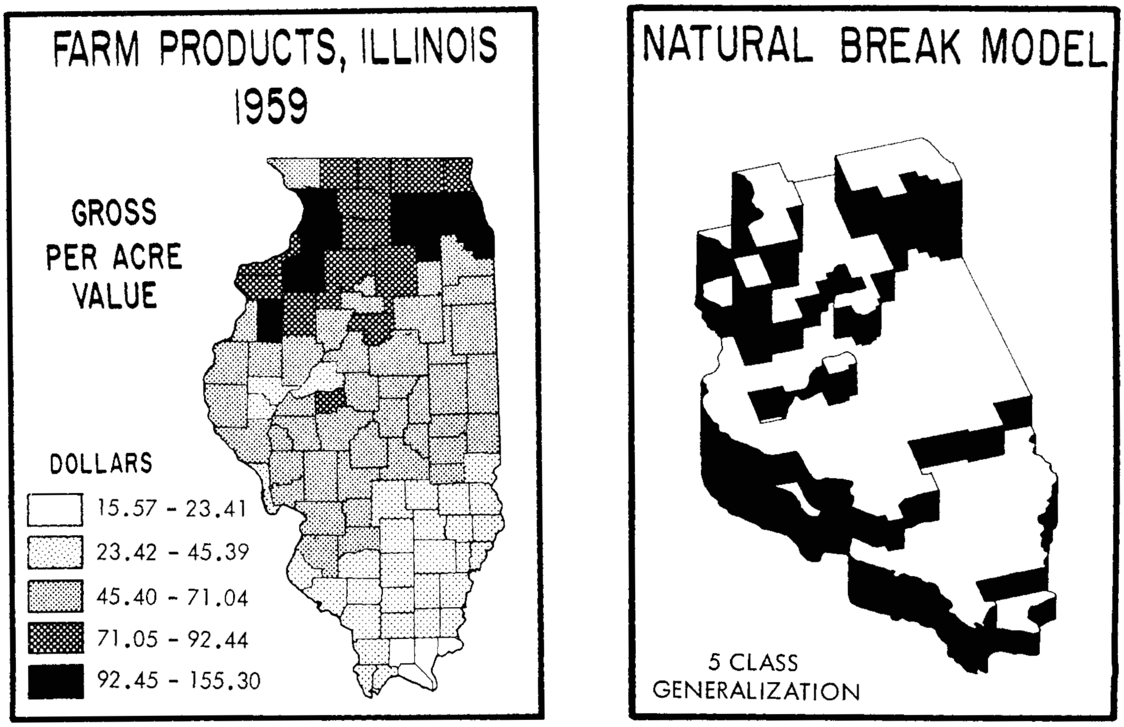


Fig. 7. This generalization of the Illinois data was prepared using the first four "natural breaks" in the histogram in Figure 6A.

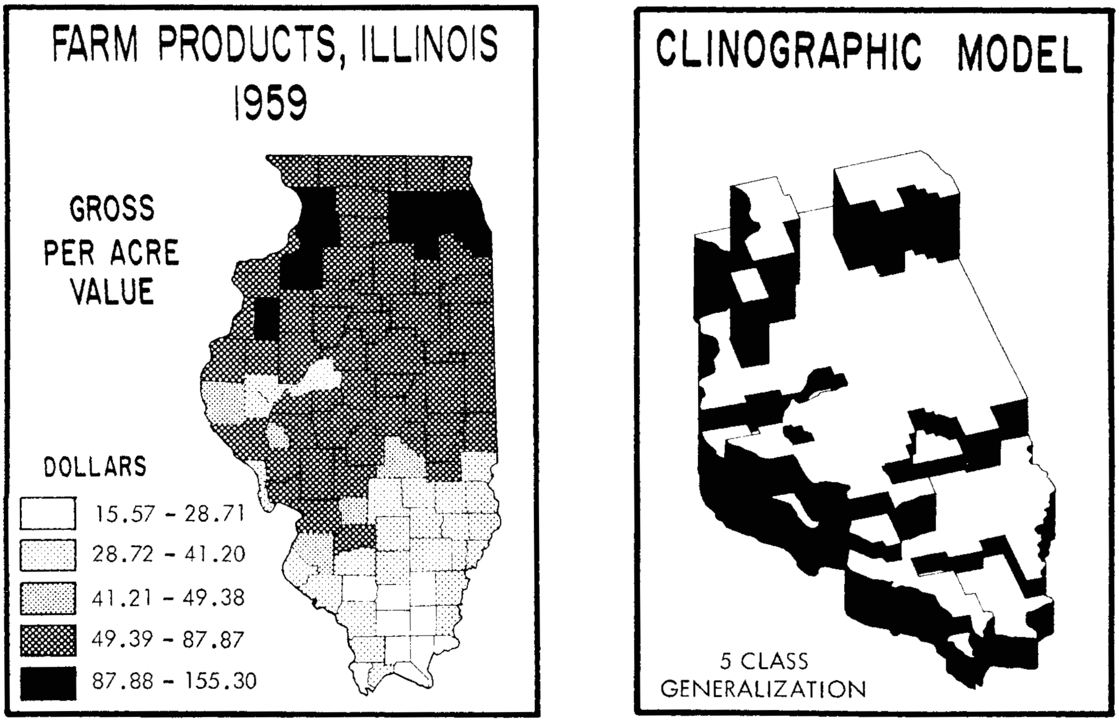


Fig. 8. Class groupings for this choroplethic map were derived through the use of a clinograph.

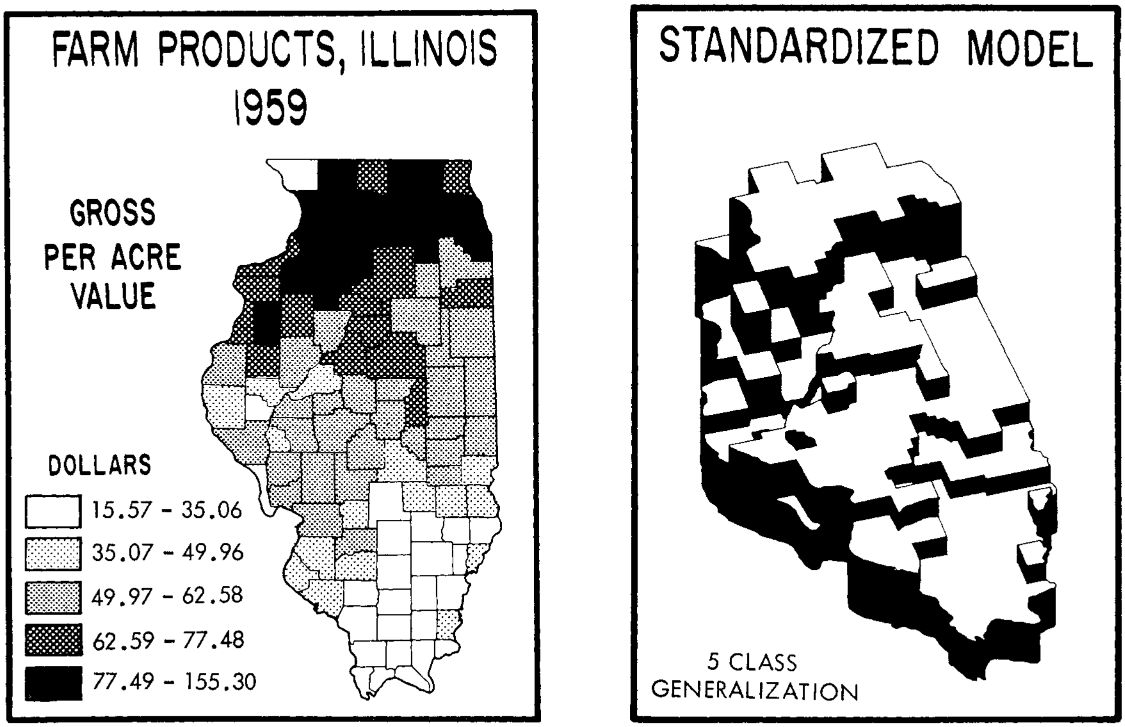


FIG. 9. Standardized map classes such as those used to create this map are obtained by dividing the statistical distribution into equal standard deviation units. Any number of classes can be determined by this procedure.

ized classes for choroplethic maps.[17] He used the standard deviation of the statistical distribution to subdivide the data into groupings (Fig. 9). Unlike other authors, Armstrong viewed the nature of the statistical distribution as paramount. The areal patterns which are a consequence of the procedure become of concern only after the map is compiled. He concluded that areal patterns created in this fashion will make choroplethic maps more meaningful because "their usefulness would be improved by adopting standard categories and by interpreting the distribution patterns with reference to the technical characteristics."[18]

Five maps with very different areal patterns of the Illinois distribution have been presented (Figs. 2, 5, 7, 8, and 9). The difference in visual impact of these maps is the result of the classification technique used to subdivide the data into classes, since other construction procedures have been held constant. All of the maps contain the same level of generalization (five classes), all are symbolized identically, and the data for each have been correctly compiled and represented. Since each map communicates a different concept of the distribution, it is quite apparent that these concepts cannot be equally valid. In these circumstances, each map-author faces a dilemma, for he can select only one map to place before his readers. How can he judge the relative worth of these representations? Must he rely on subjective judgments? If so, what criterion should be utilized in arriving at a decision as to the "best" map? These questions have led the authors to three additional queries which seem to be pertinent. They are:

1) Which map creates the most accurate overview?
2) Which map provides the reader with the most accurate intensity values for specific places?
3) Which map contains boundaries which occur along major breaks in the statistical surface?

[17] Armstrong, op. cit., footnote 5, pp. 382–90.
[18] Armstrong, op. cit., footnote 5, p. 390.

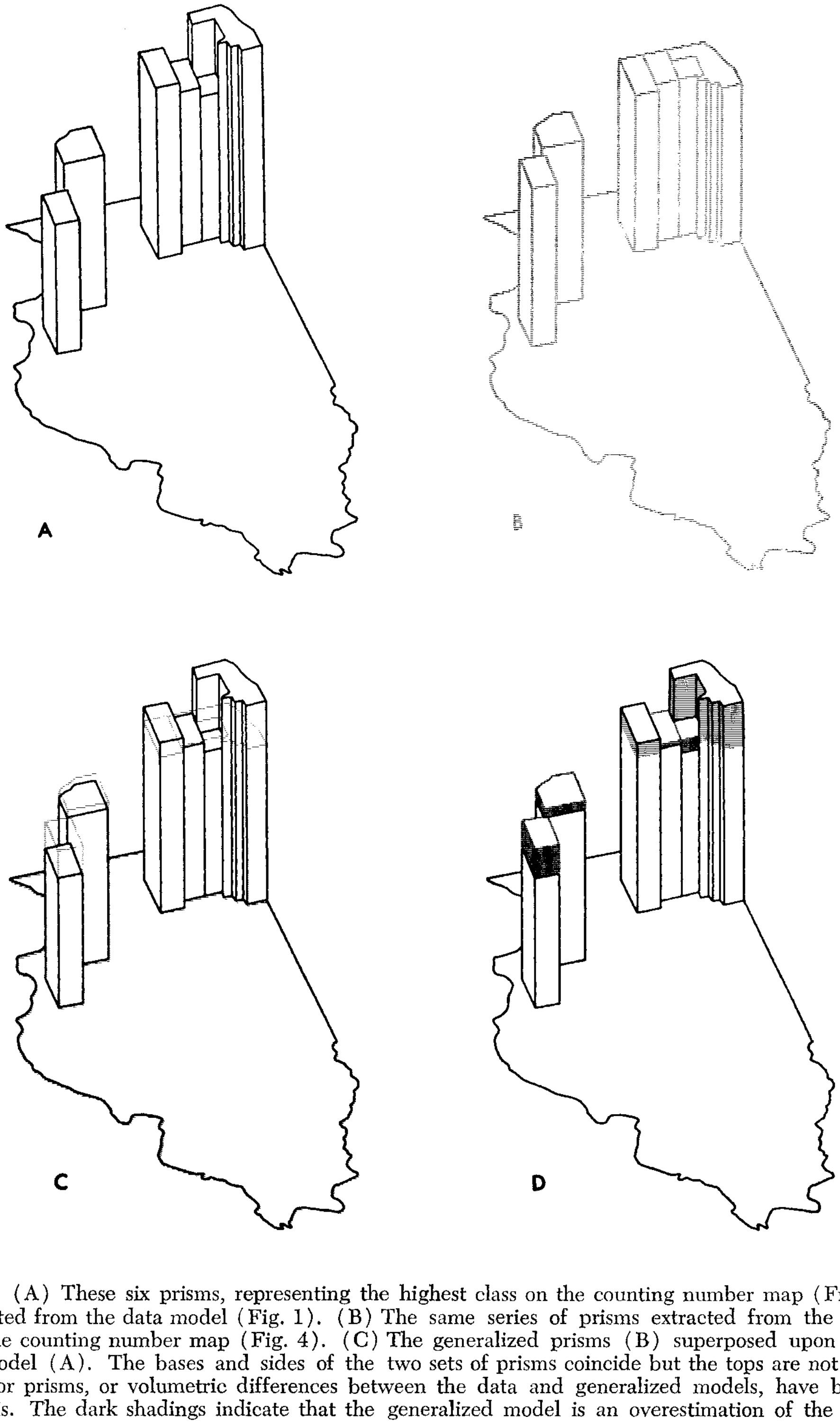


FIG. 10. (A) These six prisms, representing the highest class on the counting number map (Fig. 2), have been extracted from the data model (Fig. 1). (B) The same series of prisms extracted from the generalized model of the counting number map (Fig. 4). (C) The generalized prisms (B) superposed upon those from the data model (A). The bases and sides of the two sets of prisms coincide but the tops are not alike. (D) The six error prisms, or volumetric differences between the data and generalized models, have been shaded for emphasis. The dark shadings indicate that the generalized model is an overestimation of the data model in four enumeration units. The light shadings, on the other two enumeration units, indicate an underestimation.

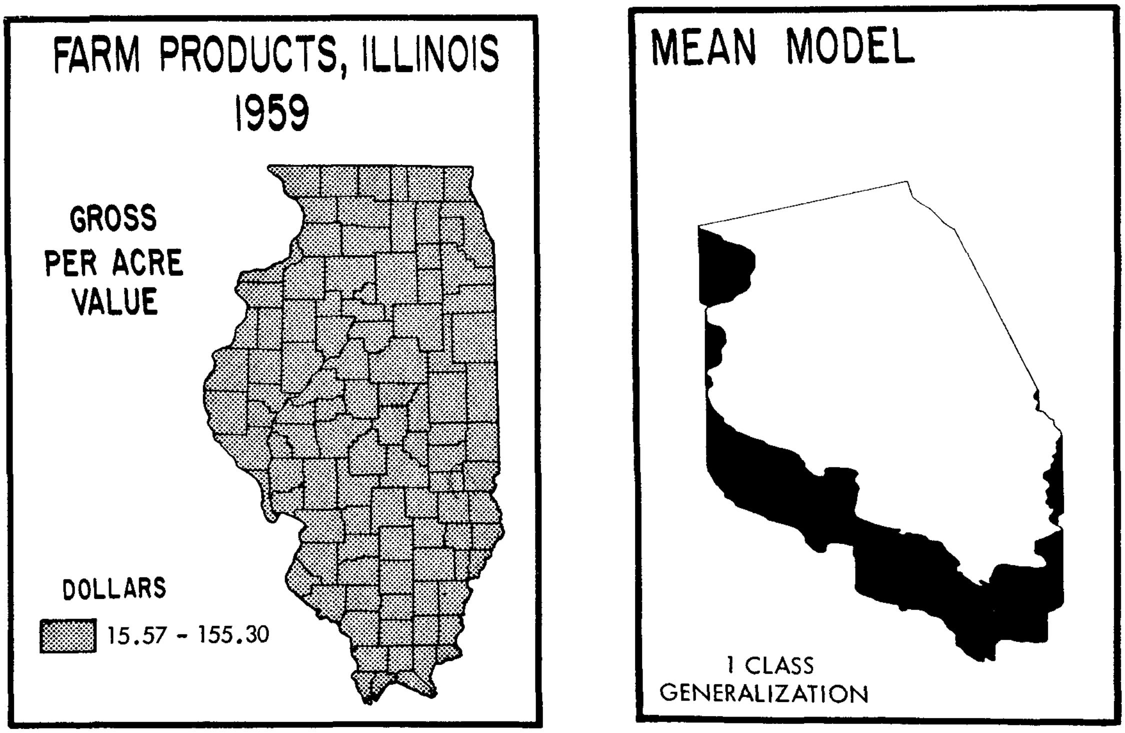


FIG. 11. This mean generalization is a useless map because it cannot communicate areal patterns. It is the grossest possible generalization of the data and is presented to illustrate one end of the error spectrum. When the mean model is compared to the data model (Fig. 1) the reader perceives the termini of the full error spectrum. All other generalizations have a surface configuration between these two models.

Finally, we have concluded that subjective analyses of such varied and complex patterns as those contained on these maps are of questionable value. An objective evaluative procedure is needed to determine the quality of a generalization.

Overview Error in Choroplethic Maps

Earlier in this paper the reality of an areal distribution was represented in the form of a data model. Later, it was shown that a choroplethic map is a planimetric representation of this data model and that it can be conceived of in three-dimensional format as a generalized model. If one accepts these concepts, it logically follows that the volumetric space between the surfaces of these two models is the overview error factor of the generalization. This holds since the most accurate overview of a distribution must be obtained from a generalized model which most closely approximates the data model.

The volume of error between a generalization and reality is composed of a series of error prisms (one for each enumeration unit) which lie between the surface of the data model and the generalized model. There are one hundred and two error prisms for the Illinois data set, and when they are placed in their respective geographic positions they form a "blanket of error" between the generalization and reality.[19] The magnitude of error for a generalization equals the absolute sum of the volume of these error prisms which comprise the blanket of error.

Six error prisms, along with the appropriate prisms from the data and generalized models, have been extracted to present this concept in a simplified illustration (Fig. 10). These six prisms are from the largest class of the counting number map (Fig. 2). One can locate their position in the total distribution by comparing Figure 11 with Figures 1, 2, 3, and 4. These six error prisms vary in height, and

[19] G. F. Jenks, "The Data Model Concept in Statistical Mapping," *International Yearbook of Cartography*, Vol. 7 (1967), pp. 186–90.

TABLE 2.—OVERVIEW ERROR INDICES

Generalizing Technique	Volume of the Error Factor	Overview Error Index (OEI)	Overview Accuracy Index (OAI)
Counting Number	342,316.61	.31026	.68974
Quintile	377,874.57	.34249	.65751
Natural Breaks	329,292.57	.29846	.70154
Clinographic	454,692.69	.41212	.58788
Standardized	347,991.72	.31541	.68459

Source: calculated by authors.

they have both positive and negative error volumes. The two positive error prisms (shown in light gray) indicate that the generalization is an underestimation of the intensities of these two enumeration units.

The absolute sum of the volumes of the error prisms is a measure of the overview quality of a generalization. When this measure is calculated for each of the five generalizations previously presented, it is possible to select the map which best suits the needs of the overview map-reader (Table 2). The natural break map should be selected for this reader because it has the smallest overview error volume. It is difficult to rate the natural break map of the Illinois distribution with that of another distribution, however, since one does not know the complete spectrum of error for this data series. To provide this statistic it is first necessary to set theoretical limits (the most accurate and the least accurate possible generalizations) of overview error for the gross farm value data. With these limits one can calculate a comparative overview measure for any level or type of generalization of these data.

The most accurate map that can be made of the Illinois data is a one hundred and one class map which would have a generalized model identical with the data model.[20] In contrast to this ungeneralized map is one which would contain but one class to represent the mean of the data (Fig. 11). The error volume between the mean model and the data model is the maximal error which can be obtained from the data; it is equal to 1,103,288.87. The theoretical error spectrum for the overview map-reader is thus limited by an accurate value of 0 and a maximum inaccurate value of 1,103,288.87. All maps of this distribution will have error volumes within this range of values, and a comparative measure of accuracy can be achieved by dividing the error volume of a generalization by the larger (inaccurate) value. The overview error index, or comparative measure, for each of the generalizations is given in the middle column of Table 2.

When the overview error index (OEI) is subtracted from one, a complementary statistic, the overview accuracy index (OAI) is obtained. This latter index, the OAI, will be used in further discussions of overview error on choropleth maps. It can range from 1.0 for a perfect or accurate map to 0 for the least accurate or mean map. Every generalization with more than one class will have an OAI with a value greater than 0 and less than 1.0.

The overview accuracy index (OAI) serves the map-maker in several ways:

1) it allows him to rank a generalization within the total error spectrum;
2) it enables him to comprehend the fact that the natural break generalization with an OAI of .70154 is only seventy percent accurate;
3) he can compare the accuracy of this map with that of another distribution in terms of overview accuracy; and
4) this OAI statistic can be combined with other error measures to be developed later so that a comprehensive error index for a generalization can be obtained.

Tabular Error in Choroplethic Maps

The tabular user of a choroplethic map typically locates an enumeration unit of interest, observes the shading pattern on it, and then obtains the intensity value range for that shading from the map legend. Obviously, he will be most satisfied when the class interval is narrow because he will obtain the most accurate estimate of the intensity values in these cases. In three-dimensional format, the tabular user is therefore more concerned with the heights of the error prisms within a class than he is with the volume of the error prisms. This means that he is primarily concerned with the vertical distances between the prisms of the data and the generalized models.

The tabular error for a single enumeration unit is the vertical distance between the plane

[20] There are one hundred and two enumeration units but two have identical intensity values. Thus, there are one hundred and one different possible classes.

TABLE 3.—TABULAR ERROR INDICES

Generalizing Technique	Total Tabular Error	Mean or Average Tabular Error	Tabular Error Index (TEI)	Tabular Accuracy Index (TAI)
Counting Number	$602.72	5.91	.31199	.68801
Quintile	$645.25	6.33	.33400	.66600
Natural Breaks	$569.42	5.58	.29475	.70525
Clinographic	$773.18	7.63	.40022	.59978
Standardized	$595.52	5.84	.30826	.69174

Source: calculated by authors.

of that unit on the data and generalized models. The total tabular error for a generalization is the absolute sum of these vertical distances for all enumeration units in the mapped area. The total tabular error for the five generalizations is given in the first column of Table 3. When the total tabular error is divided by the number of enumeration units the mean tabular error is obtained (column two, Table 3). The mean tabular error is a particularily meaningful measure for the tabular map-reader since it enables him to understand the level of accuracy he can expect on the map.

As in the case of the overview error volumes, the tabular error is not particularily helpful until it can be put within the total tabular error spectrum. The minimum tabular error is obtained when the planes of the data and generalized models coincide, or when the data are presented on a one hundred and one class map. Similarly, the maximal tabular error is obtained from a one class (the mean) map, in which case a sum of $1931.84 is obtained. Using these tabular error limits, it is possible to compute tabular error indices (TEI) and tabular accuracy indices (TAI) for the five generalizations (columns three and four, Table 3). The natural break map is the most accurate map for the tabular user, since it has the highest TAI.

The tabular and overview accuracy indices are closely related, since the first measures the thickness of the blanket of error and the latter the volume of this error complex. Since this is the case, the map-maker need not calculate the OAI if his data were collected on a uniform areal grid. Conversely, the OAI becomes most important when there is a significant variation in the areas of enumeration units.

Boundary Error in Choroplethic Mapping

The boundaries between shadings on a choroplethic map tend to dominate the visual impact of the representation, because sharp visual contrasts occur along these lines. Map-readers tend to assign significance to these boundaries and, as a result, often assume that they designate breaks in the configuration of the statistical surface. Since this seems to be the normal reaction among map-users, the map-maker is obliged to use generalizations in which there is a concurrence of boundaries and surface breaks.

Two hundred and fifty eight boundary segments separate the one hundred and two counties of Illinois. These boundary segments are the black "cliffs" on the data model; they vary in height from $0 to $92.89 (Fig. 1). The numerical value for each "cliff" is the difference between the intensity values of the two counties which share a given boundary segment. The statistical distribution of these two hundred and fifty eight values is presented on the histogram in Figure 12A.

Since each generalization is a simplification of reality, some of the "cliffs" on the data model surface become boundary "cliffs" and the others are disregarded. The boundary values used to delineate regions on the generalized model of the quintile map have been shown in black in Figure 12B, and the within-region boundaries, or the unused "cliffs," are shown in gray. In this case one hundred and nine boundary values (the black dots) were used in the generalization and one hundred and forty nine "cliffs" (the gray dots) were unused.

Ideally, the boundary values utilized in a generalization should coincide with the highest "cliffs" on the data model. Assuming that one hundred and nine cliffs (this number was used in the quintile map) are needed to create a generalization, the ideal set of one hundred and nine boundaries would be that shown in black in Figure 12C. The values of the bounding cliffs used for the quintile generalization (black dots in 12B) total $1624.08. When this value is divided by $2401.54, the sum of the ideal one hundred and nine boundary values (black dots in Figure 12C), a measure of boundary accuracy of .67626 is obtained. This boundary accuracy index (BAI) measures the efficiency with which a generalization achieves the ideal boundary representation. Like the other indices, it can range from

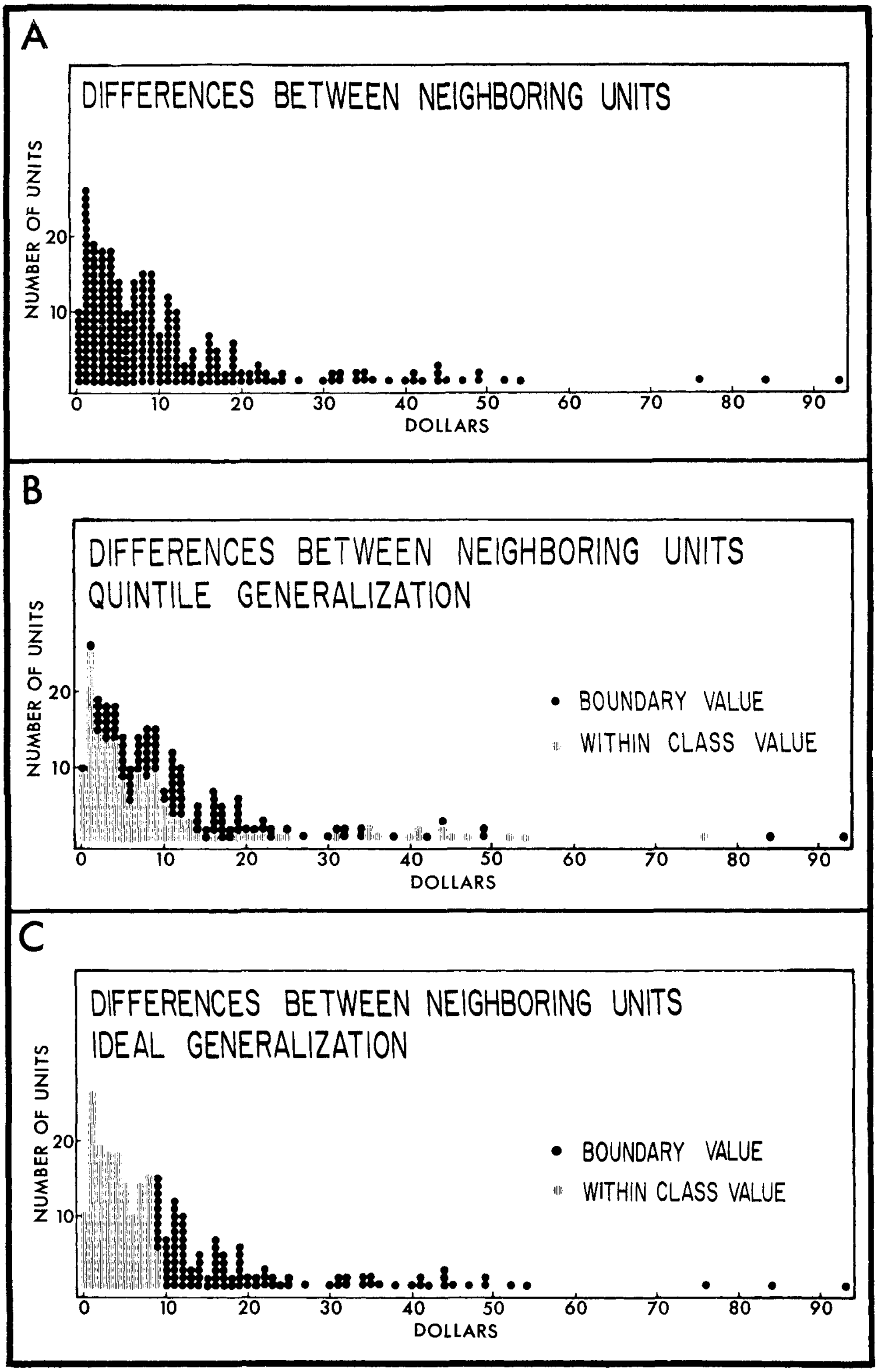

A
DIFFERENCES BETWEEN NEIGHBORING UNITS
NUMBER OF UNITS
10
20
0
10
20
30
40
50
60
70
80
90
DOLLARS
B
DIFFERENCES BETWEEN NEIGHBORING UNITS
QUINTILE GENERALIZATION
● BOUNDARY VALUE
▥ WITHIN CLASS VALUE
NUMBER OF UNITS
10
20
0
10
20
30
40
50
60
70
80
90
DOLLARS
C
DIFFERENCES BETWEEN NEIGHBORING UNITS
IDEAL GENERALIZATION
● BOUNDARY VALUE
▥ WITHIN CLASS VALUE
NUMBER OF UNITS
10
20
0
10
20
30
40
50
60
70
80
90
DOLLARS

TABLE 4.—BOUNDARY ACCURACY INDICES

Generalizing Technique	Boundary Accuracy Index (BAI)
Counting Number	.8691
Quintile	.6763
Natural Breaks	.9134
Clinographic	.8203
Standardized	.7703

Source: calculated by authors.

0 for an inaccurate map to 1 for a highly accurate representation. The boundary indices for the five generalizations shown previously are given in Table 4.[21]

Five generalizations of the Illinois data have been created and tested. In each case the natural break generalization has been the most accurate, and there is little question about using it in preference to the others. The decision in this case is clear-cut, but accuracy index values can vary from generalization to generalization. For example, the other maps are not ranked in the same order of accuracy from one measure to another (Tables 2, 3, and 4). It is therefore necessary that provision be made to select one map from a group in which no single map achieves dominance in all accuracy measures. If a map-author could anticipate his map-readers' intentions such a method would not be needed, but all too often he has no knowledge of whom his readers may be, nor how they will utilize his product. For these reasons, a single map accuracy index is both necessary and desirable.

A Composite Accuracy Index

A map-maker can view map-readers as three opposing forces, each tugging from a different direction. The overview reader wants one map attribute, whereas the tabular and boundary users want others, and the map-maker cannot always provide for each group equally well.

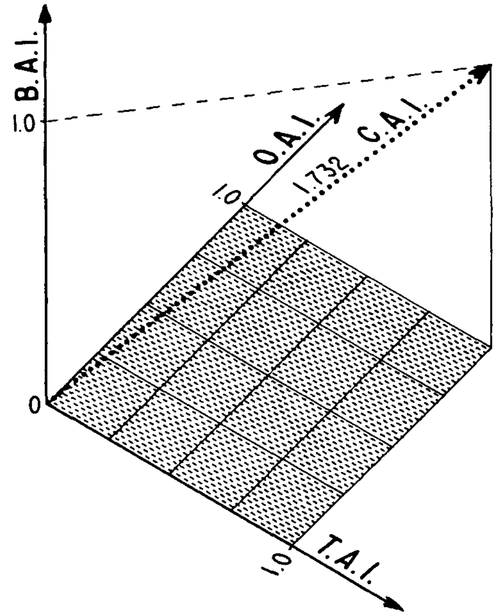


FIG. 13. The composite map accuracy index (CAI) is obtained by solving the equation: CAI = OAI^2 + TAI^2 + BAI^2.

This realization forces him to seek a compromise generalization which does the best job possible for readers of all types.

A solution to the cartographic dilemma of selecting a compromise generalization is found in vector algebra (Fig. 13). If the one hundred and one class map is taken as an example, the OAI, TAI, and BAI values for this map are identical, and all have the numerical value of 1.0. This perfect, although impractical, map would have a "composite" accuracy index as follows:

$$\begin{aligned} \text{CAI} &= (\text{OAI}^2 + \text{TAI}^2 + \text{BAI}^2)^{\frac{1}{2}} \\ &= (1 + 1 + 1)^{\frac{1}{2}} \\ &= (3)^{\frac{1}{2}} \\ &= 1.732 \end{aligned}$$

The composite index for a generalization must be less than 1.732, because there is always

[21] The authors are indebted to C. Gregory Knight for his assistance in developing this boundary measure.

←

FIG. 12. (A) The statistical distribution of the boundary differences ("cliffs") on the surface of the gross farm products distribution (Fig. 1). (B) Some of the boundary differences of "cliffs" are used as bounding lines for shadings on each generalization. The 109 used on the quintile generalization are shown as black dots. The unused "cliffs" are shown in gray. (C) Ideally a generalization would contain boundaries which coincided with the greatest "cliffs." The 109 greatest "cliffs" are shown in black on this graph. The boundary accuracy of a representation is the ratio of those actually used (B) to the same number of the greatest boundaries that could have been used (C). The sum of the 109 boundaries used on the quintile generalization is $1,624.08, the sum of the 109 greatest is $2,401.54, and the ratio between these sums is .67626.

TABLE 5.—MAP ACCURACY INDICES

Generalizing Technique	OAI	TAI	BAI	MAI
Counting Number	.68974	.68801	.86907	.75375
Quintile	.65751	.66600	.67626	.66664
Natural Breaks	.70154	.70525	.91345	.77974
Clinographic	.58788	.59978	.82026	.67779
Standardized	.68459	.69174	.77032	.71661

Source: calculated by authors.

TABLE 6.—THE NUMBER OF POSSIBLE CHOROPLETHIC GENERALIZATIONS OF THE ILLINOIS DATA

Number of Classes	Possible Number of Maps
1	1
2	101
3	5,050
4	166,650
5	4,082,925
6	79,208,745
7	1,267,339,920

Source: calculated by authors.

some error in the generalizing process. Thus for the natural break generalization the CAI is:

$$\begin{aligned}\mathrm{CAI} &= [(.7015)^2 + (.7052)^2 + (.9135)^2]^{\frac{1}{2}}\\ &= (1.8239)^{\frac{1}{2}}\\ &= 1.3504\end{aligned}$$

It is difficult to comprehend the relationship of these CAI indices with those obtained for each of the accuracy indices, however, and this is overcome by reducing the CAI to a simple proportion as follows:

$$\frac{1.3504}{1.732} = .7797$$

The proportional composite index is the composite map accuracy index, and it will be referred to as the MAI. The complete set of indices for the five generalizations are shown in Table 5.[22]

A NEEDLE IN A HAYSTACK

A map-author is in much the same position as a person looking for a needle in a haystack, because so many different representations of an areal distribution can be made. There are 4,082,925 different five-class maps which can be made of the Illinois data (Table 6). Five of these maps have been created by the traditional generalizing processes. One wonders whether any of the four million plus remaining maps would yield higher OAI, TAI, BAI, and MAI indices than those obtained. Assuming that such is the case, a question arises as to the methods by which this more accurate map can be found.

The computer has given cartographers a valuable assist in attacking generalization problems and the authors contemplated using it to create and evaluate *all* class combinations of the Illinois distribution. Table 6, however, is convincing evidence that such a "brute force" solution to the generalization problem is economically impractical. Many of the possible generalizations, at any class level, are so inaccurate that creating and evaluating them is of limited value. The poorest five-class generalization that we could invent occupies the undesirable end of the error spectrum (Fig. 14), and it is shown only to emphasize the need for some generalizing hypotheses to permit the creation of maps at the opposite end of the error scale.

The Search for High TAI Values

The tabular error for a choroplethic generalization is the sum of the vertical differences between the data and generalized (mean) prisms for all enumeration units. Since each difference value is calculated independently, the intensity and mean values can be taken out of their geographic context and arrayed along a number line. Map classes are derived by cutting this number line. When traditional generalizing procedures are followed, the cuts in the number line derive from external considerations. This defect in data handling is eliminated when a reiterative and "force" grouping procedure is used to minimize the differences around class means for the total number line. This procedure can best be described in a series of manipulations which can be grouped

[22] The TAI and OAI indices are not independent and thus the technique for calculating the MAI should not contain these values as independent vectors at right angles to each other. The problem is, however, that the relationship between the TAI and OAI varies from area to area, being dependent upon uniformity of enumeration unit areas. Further study of this problem may indicate that the OAI, TAI, and BAI indices should be combined in a different manner, e.g. by multiplication, or that the map author should select that error factor which he deems most significant. This latter solution, however, implies that the author knows who his readers will be and what type of information they will want to obtain from the map.

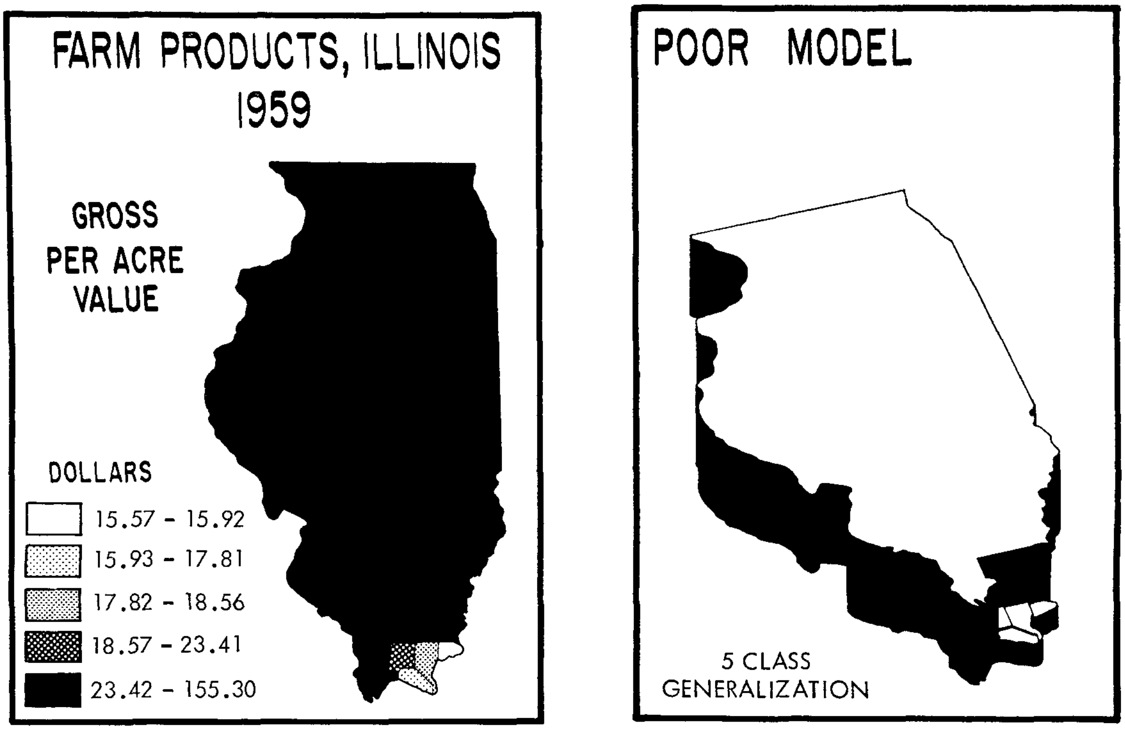


FIG. 14. This map occupies a position near the undesirable end of the spectrum of error for generalizations of the Illinois data. It is the poorest five-class example that we could conceive. The index values are: OAI = .12157, TAI = .07062, BAI = .15357, and MAI = .12017.

into two steps. In addition to the description which follows, a practical example is detailed in Table 7.

Step One. Reiterative Cycling

a) The arrayed data are cut into arbitrary classes.

b) Means for these classes are calculated.

c) OAI, TAI, BAI, and MAI indices for the classes are calculated and stored in four separate locations along with the limits of each class.

d) A new set of classes is created by grouping the intensity values to the nearest class mean.

e) Means and accuracy indices for the new classes are calculated.

f) Each new index is compared with its stored counterpart. If it is higher, the old set of classes is destroyed and the new set is stored in its place. If the new index is lower than the stored index, the new set is destroyed.

g) The process is continued until the class limits, means, and indices repeat themselves.

Step Two. Force Cycling

a) Starting with the classes obtained above, a new set is created by forcing one member into the class with the smallest mean, taking it from the class with the next smallest mean. (Remember that the data are arrayed and thus the smallest intensity in the second class is moved to the first class.)

b) Calculate means and indices, and compare as before.

c) If the TAI is increased move a second member into the class with the smallest mean. If the TAI is decreased discard the first set of force classes and return to the last set derived by reiteration.

d) Force a member from the class with the third largest mean into the class with the second largest mean. Calculate and compare as above (2b and c). Continue if TAI is increased; if not, go to 2e.

ARBITRARY CLASSES

INTENSITY VALUES	CLASS MEAN	SUM OF DIFFERENCES
4		
5		
6		
6	7.250	16.000
8		
9		
10		
10		
11		
11		
12	12.833	9.000
13		
14		
16		
17		
18		
18		
18	18.833	8.666
20		
22		Total 33.666

REITERATION 1

INTENSITY VALUES	CLASS MEAN	SUM OF DIFFERENCES
4		
5		
6		
6	7.250	16.000
8		
9		
10		
10		
11		
11		
12	12.200	5.200
13		
14		
16 *		
17		
18		
18		
18	18.428	10.284
20		
22		Total 31.484

REITERATION 2

INTENSITY VALUES	CLASS MEAN	SUM OF DIFFERENCES
4		
5		
6		
6	6.333	8.666
8		
9		
10 *		
10 *		
11		
11		
12	11.571	8.571
13		
14		
16		
17		
18		
18		
18	18.428	10.284
20		
22		Total 27.521

REITERATION 3

INTENSITY VALUES	CLASS MEAN	SUM OF DIFFERENCES
4		
5		
6		
6	5.800	5.200
8		
9 *		
10		
10		
11		
11		
12	11.250	10.500
13		
14		
16		
17		
18		
18		
18	18.428	10.284
20		
22		Total 25.984

REITERATION 4

INTENSITY VALUES	CLASS MEAN	SUM OF DIFFERENCES
4		
5		
6		
6	5.800	5.200
8		
9		
10		
10		
11		
11		
12	11.250	10.500
13		
14		
16		
17		
18		
18		
18	18.428	10.284
20		
22		Total 25.984

FORCE 1

INTENSITY VALUES	CLASS MEAN	SUM OF DIFFERENCES
4		
5		
6		
6	6.333	8.666
8		
9 *		
10		
10		
11		
11		
12	11.571	8.571
13		
14		
16		
17		
18		
18		
18	18.428	10.284
20		
22		Total 27.521

FORCE 2

INTENSITY VALUES	CLASS MEAN	SUM OF DIFFERENCES
4		
5		
6		
6	5.800	5.200
8		
9		
10		
10		
11		
11		
12	11.777	15.777
13		
14		
16 *		
17		
18		
18		
18	18.833	8.666
20		
22		Total 29.643

FORCE 3

INTENSITY VALUES	CLASS MEAN	SUM OF DIFFERENCES
4		
5		
6		
6	5.800	5.200
8		
9		
10		
10		
11		
11		
12	10.857	7.143
13		
14 *		
16		
17		
18		
18		
18	17.875	13.250
20		
22		Total 25.593

FORCE 4

INTENSITY VALUES	CLASS MEAN	SUM OF DIFFERENCES
4		
5		
6		
6	5.800	5.200
8		
9		
10		
10		
11		
11		
12	10.500	5.000
13 *		
14		
16		
17		
18		
18		
18	17.333	18.667
20		
22		Total 28.867

FORCE 5

INTENSITY VALUES	CLASS MEAN	SUM OF DIFFERENCES
4		
5		
6		
6	5.250	3.000
8*		
9		
10		
10		
11		
11		
12	10.500	10.000
13		
14		
16		
17		
18		
18		
18	17.875	13.250
20		
22		Total 26.250

FORCE 6

INTENSITY VALUES	CLASS MEAN	SUM OF DIFFERENCES
4		
5		
6		
6	5.800	5.200
8		
9		
10		
10		
11		
11		
12	10.500	5.000
13*		
14		
16		
17		
18		
18		
18	17.333	18.667
20		
22		Total 28.867

FORCE 7

INTENSITY VALUES	CLASS MEAN	SUM OF DIFFERENCES
4		
5		
6		
6	5.250	3.000
8*		
9		
10		
10		
11		
11		
12	10.500	10.000
13		
•		
14		
16		
17		
18		
18		
18	17.875	13.250
20		
22		Total 26.250

FORCE 8

INTENSITY VALUES	CLASS MEAN	SUM OF DIFFERENCES
4		
5		
6		
6	6.333	8.666
8		
9*		
10		
10		
11		
11		
12	11.166	5.332
13		
14		
16		
17		
18		
18		
18	17.875	13.250
20		
22		Total 27.248

FORCE 9

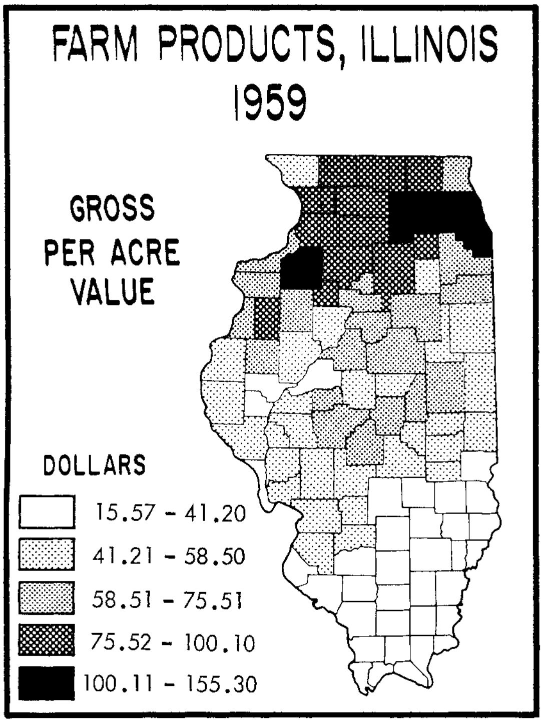
INTENSITY VALUES	CLASS MEAN	SUM OF DIFFERENCES
4		
5		
6		
6	5.800	5.200
8		
9		
10		
10		
11		
11		
12	11.250	10.500
13		
14*		
16		
17		
18		
18		
18	18.428	10.284
20		
22		Total 25.984

REITERATION 5

INTENSITY VALUES	CLASS MEAN	SUM OF DIFFERENCES
4		
5		
6		
6	5.800	5.200
8		
9		
10		
10		
11		
11		
12	10.857	7.143
13		
14*		
16		
17		
18		
18		
18	17.875	13.250
20		
22		Total 25.593

TABLE 7

This table is designed to illustrate the reiteration and forcing method of grouping mappable data. The starred numbers indicate intensity values that were moved from one class to another. The following listing gives the rationale for the grouping. Arbitrary Classes; Reiteration 1, accept new classes because Total Sum of Differences is reduced; Reiteration 2, accept; Reiteration 3, accept; Reiteration 4 repeats 3 so cycle is temporarily terminated; start force cycle; Force 1, reject, return to Reiteration 3 classes; Force 2, reject, return to R3 classes; Force 3, accept; Force 4, reject, return to F3 classes; Force 5, reject, go to F3; Force 6, reject, go to F3; Force 7, reject; Force 8, reject; Force 9, reject, forcing cycle completed go to reiteration. Reiteration 5, F3 classes repeat themselves, terminate procedure and accept F3 classes because the minimal sum of deviation was achieved here.



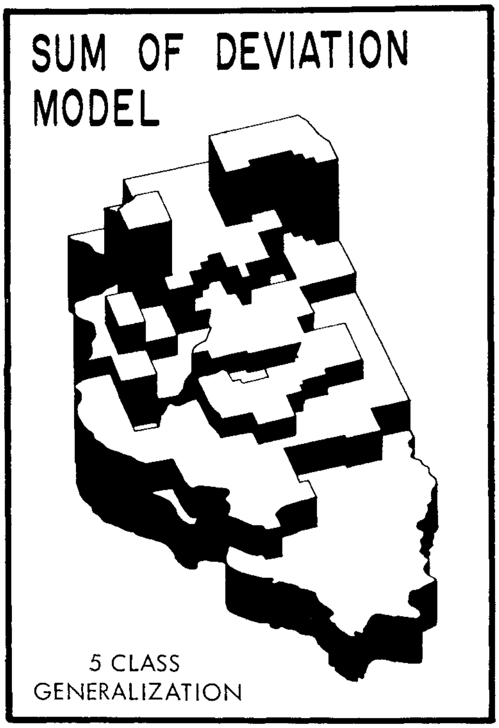


Fig. 15. One of the best, if not the best map of the Illinois data for the tabular map-user. The TAI for this map is .73455, which is nearly three percentage points higher than any of the "traditional" maps (Table 5). This map is also the best representation for the overview map-user, with an OAI of .74842.

e) Forcing is continued until the class with the highest mean is reached. At this point upward forcing is initiated by moving a member from the class with the second largest mean into the class with the largest mean. Test and compare as before. This procedure is continued until the class with the smallest mean is reached.

f) Continue the procedure with another downward forcing cycle.

g) Continue the procedure with another upward forcing cycle.

h) Continue the procedure by duplicating the set of classes with the highest TAI index (in storage) and reiterate it as in Step One.

At the conclusion of this two step procedure the TAI storage contains a set of map classes which can be characterized as *one of the best*, ,if not *the best*, attainable for the tabular map-user. This may appear to be a strong statement, but we have been unable to generate, either purposefully or by accident, a better tabular representation in any set of data that we have processed in this manner (Fig. 15). The TAI of .73455 for this map is nearly three percentage points higher than any TAI achieved previously (Table 5).

The Search for High OAI Values

Overview error, like tabular error, is derived on an individual basis in that error volume is calculated independently for each enumeration unit and then error volumes are summed. The reiterative and forcing program used for the TAI search can again be utilized but with a slight modification, which involves calculating an error volume to replace the linear error for each member of a class. The OAI of each set of classes derived with these error volumes is compared with that stored from the previous (TAI) procedures.

The best, or at least one of the best, representations for the overview map-user was created in the TAI manipulations; it is identical with the best TAI map (Fig. 15). The OAI attained by this set of classes is .74842, or more than three percentage points higher than any achieved by the traditional groupings.

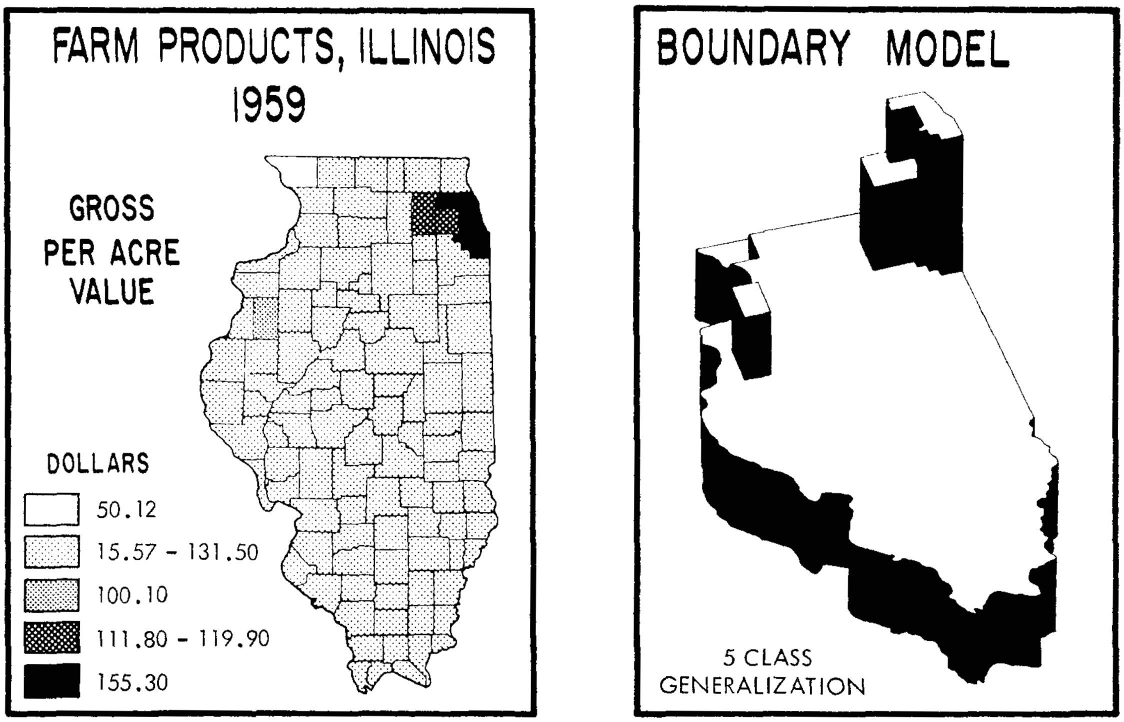


FIG. 16. This boundary map and model were created by utilizing the largest boundary difference values to delineate regions or classes. Each of the classes formed by this process is distinctly different from its neighbor. See the data model (Fig. 1). The overlapping values in the legend confuse the map-reader, however, since he cannot understand why some enumeration units have been selected for special treatment. Why, for example, should the enumeration unit with an intensity of 50.12 be in a separate class when other similar intensity values are grouped together?

One should not be surprised that the best OAI and TAI representations of the Illinois data are identical, since the areas of the counties in Illinois are not extremely varied. It would be possible to obtain similar results, even with greatly divergent areal sizes, if the intensity value of the large enumeration units were similar in value to one or more of the class means.[23]

The Search for Higher BAI Values

In the search for high TAI's and OAI's the data (intensity values) were taken out of geographic context and manipulated along a number line. This was possible because error values were calculated enumeration unit by enumeration unit, and then summed at the end of each cycle of manipulation. This cannot be done in the boundary accuracy search, because boundary differences occur between neighboring enumeration units, and their value depends upon the relative intensity values of these neighbors. Further, an enumeration unit may have several neighbors, and thus several separate difference values may be associated with it. For example, there are 258 boundary differences ("cliffs") on the Illinois data surface, and since there are 102 enumeration units

[23] The optimal solution for the Overview Accuracy Index (OAI) is achieved by the maximization of the objective function

$$1 - \frac{\sum_{j=1}^{m}\sum_{k=1}^{n} \left| Z_{kj} - \bar{Z}_j \right| A_{kj}}{\sum_{i=1}^{n} \left| Z_i - \bar{Z} \right| A_i},$$

where,

Z_i is the i^{th} element of a vector of n observations of a variable,

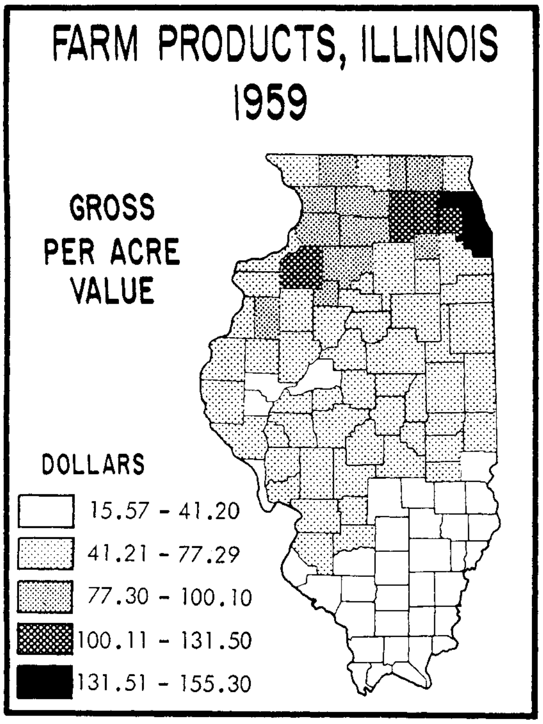
A_i is the area of the data collection unit corresponding to the i^{th} element of the vector of n observations,

$\bar{Z}$ is the mean of the n observations,

Z_{kj} is the k^{th} element of the j^{th} class which contains n_j observations of the variable Z,

A_{kj} is the area of the data collection unit corresponding to the k^{th} element in the j^{th} class, and

$\bar{Z}_j$ is the mean of the n_j observations in the j^{th} class.



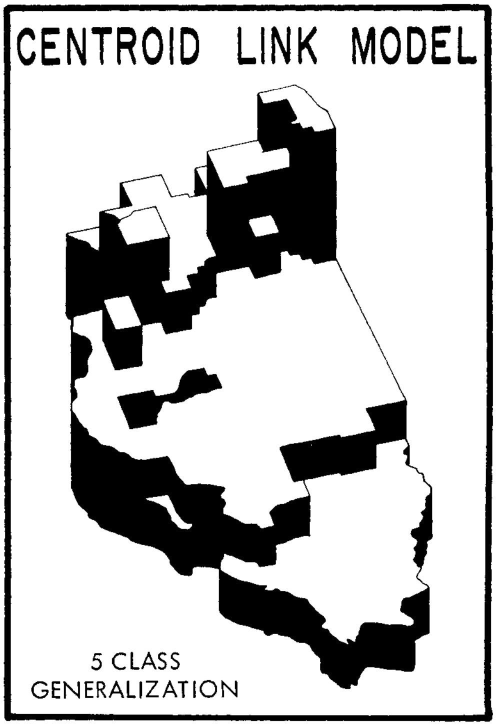


Fig. 17. In centroid linkage, most-similar pairs are found and the intensity values for these individuals are replaced by their mean intensity. Grouping is accomplished by adding the member which is closest to the mean and recalculating a mean for the new group. This proceeds until the desired number of groups are attained.

there are, on the average, two or three neighbors for each unit.

The forcing technique for minimizing choroplethic map error is based upon the ability to create new generalizations by moving one intensity value in each manipulative cycle. After such a move the new classes are evaluated and the accuracy index is compared with the one attained in an earlier cycle. An intensity value (enumeration unit value) must be moved from one class to another, and several boundary differences are involved when such a move takes place. Since boundary differences are not related to the intensity value being moved, but are controlled by the relationship of that value to all neighboring intensity values, the problem of identifying the most promising enumeration unit maneuver is not readily soluble. Unable to resolve this grouping dilemma by using the reiterative and forcing technique, we have temporized by attempting to use alternative procedures.

Since the BAI measures the degree to which boundaries on the map (or generalized model) coincide with breaks on the data surface, one can logically assume that high boundary differences should be utilized to delineate choroplethic map regions. A map was created by first drawing in the boundary with the largest difference, then the second largest difference, and continuing until five regions were outlined. Comparison of this model (Fig. 16) with the data model (Fig. 1) clearly shows that many of the major breaks in the data surface were utilized to advantage in this generalization. The BAI for this map, however, was only .82614, because some of the major boundary breaks in the surface do not lie in contiguous positions and therefore were not used in delineating the regions. These large, but unconnected and therefore unused, boundary differences became within-class boundaries and thus reduced the boundary accuracy ratio. The boundary accuracy rating for this map is less than that attained by the counting number and natural break maps created by the traditional procedure (Table 5).

Not only is the "Boundary Map" less accu-

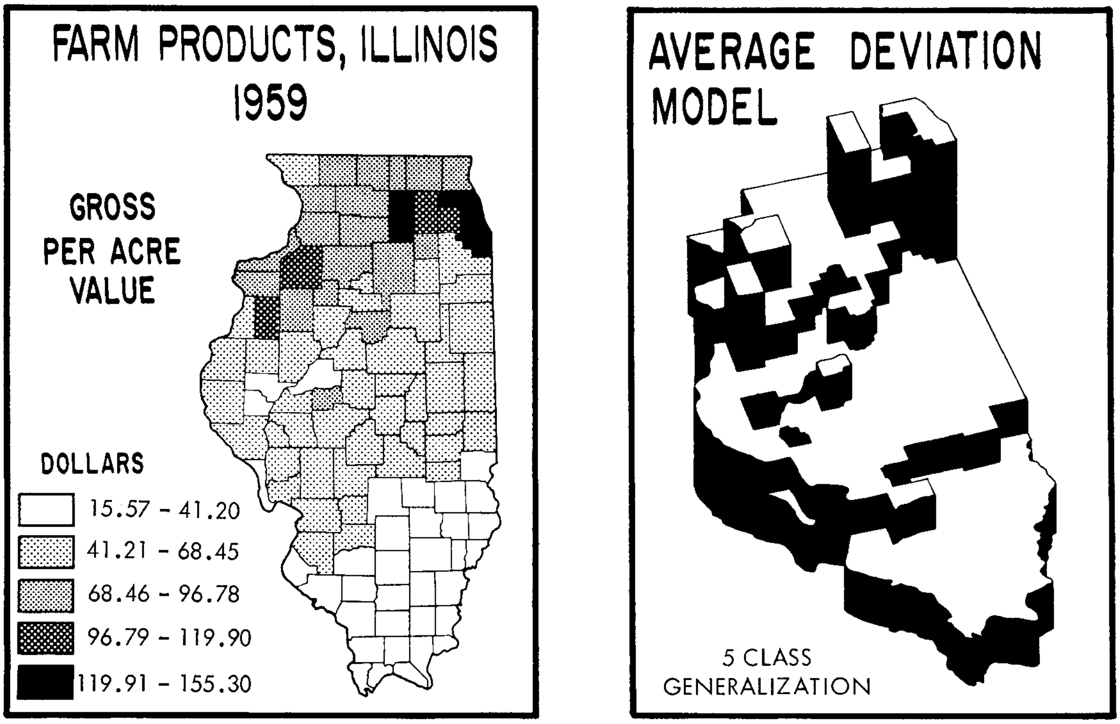


FIG. 18. The average deviations about the means of each class were equalized in the grouping procedure used to create this map. The high BAI value of .94249 indicates that this method of generalization produces maps which are useful to the boundary map-reader.

rate than had been anticipated, but it has several qualities which make it a very poor generalization. First, and most significant, the technique delineated overlapping classes. One class (15.57–131.50) encompasses within its intensity range all of the values of three other classes. Although these encompassed classes represent areal units which differ from their neighbors on the data surface, the reader of the final map (Fig. 16) is not necessarily apprised of this fact. He can only wonder why these areas were selected for special status, when similar values seem to be included in another class. All in all, the boundary difference procedure for generalizing intangible distributions was judged to be inadequate and other techniques were considered.

Geographers believe that the distributions that they study, and map, derive from some ordering. Following this tenet, one may disallow the concept of randomness and look for some system of grouping his observations which will gather like things or values into a single class. The linkage techniques of classification become manipulative modes for carrying out this rationale. Three linkage techniques (single, multiple, and centroid) were applied to the Illinois data, and one of the resultant maps and models is presented in Figure 17.[24] The boundary accuracy indices attained by these manipulative modes are given in Table 8. The centroid linkage mode attained a BAI of .93184, which surpassed any of those achieved previously.

Two additional grouping techniques which are related to the reiterative-forcing procedures were also included in the search for a high BAI value. These procedures force not on low TAI, OAI, or BAI indices, but on average class deviations or on minimum blanket-of-error measures for the classes.[25] In the first,

[24] The linkage techniques used in this paper are discussed more fully in R. R. Sokal and P. H. A. Sneath, *Principles of Numerical Taxonomy* (San Francisco: W. H. Freeman & Co., 1963), and R. J. Johnston, "Choice in Classification: The Subjectivity of Objective Methods," *Annals*, Association of American Geographers, Vol. 58 (1968), pp. 575–89.

[25] Jenks, op. cit., footnote 19, Figure 10 following page 186.

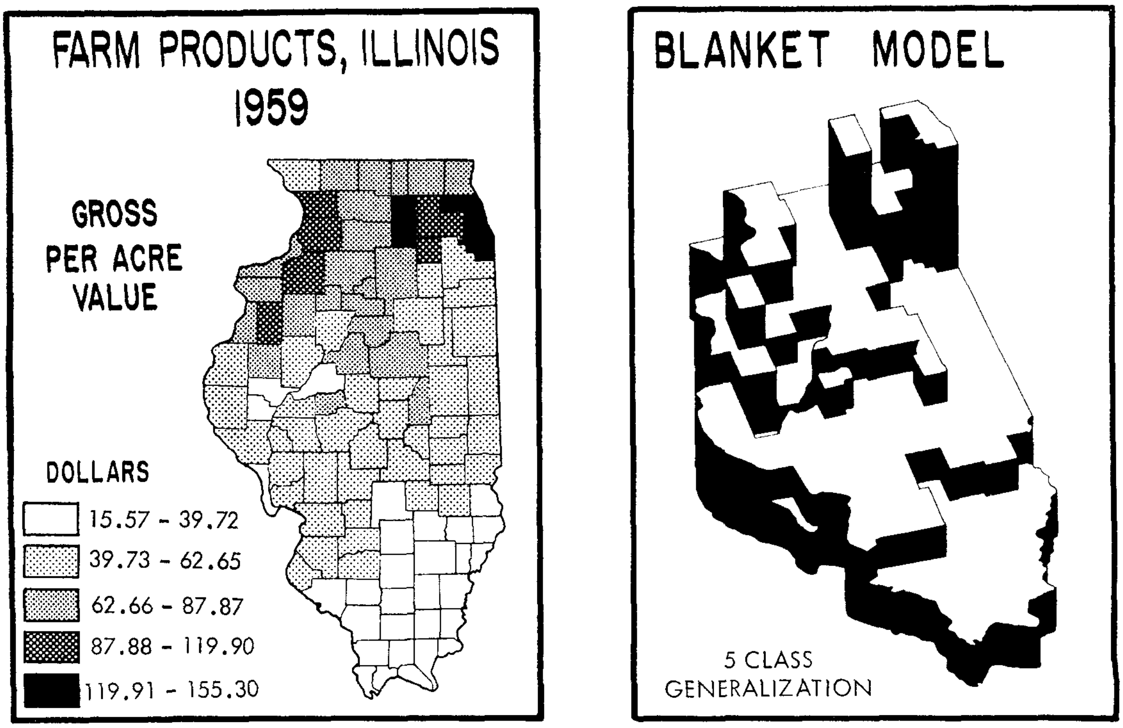


FIG. 19. The maximum error for each class has been equalized in this generalizing process. This technique produces high BAI values in many levels of generalization and for varied types of data sets.

arbitrary classes are created and means are then calculated. Differences from these means are obtained for each class and the mean of these differences, or the average deviation, is obtained. Forcing moves are selected so that all classes attain average deviations which are equal, or as nearly equal as is possible. After each forced move, TAI, BAI, OAI, and MAI indices are calculated and compared. The map resulting from this grouping process is shown in Figure 18. The BAI of .94249 indicates that this procedure is capable of producing good representations for the boundary map-user.

TABLE 8.—ACCURACY INDICES FOR MAPS CREATED BY NEW TECHNIQUES

Generalizing Technique	OAI	TAI	BAI	MAI
Poor	.12157	.07062	.15358	.12017
Boundary	.23507	.14959	.82614	.50337
Single Linkage	.34057	.26887	.79117	.52097
Multiple Linkage	.61102	.61015	.86750	.70669
Centroid Linkage	.68849	.66425	.93184	.77106
Reiteration	.71648	.69590	.94884	.79540
Blanket	.70893	.69334	.91560	.77925
Average Deviation	.71596	.69114	.94249	.79133
Sum Deviation	.74842	.73455	.89747	.79691

Source: calculated by authors.

The minimum blanket of error technique is very similar to the average deviation forcing procedure, but the forcing moves are determined by the size of the largest linear error in each class. An attempt was made to equalize this maximum error in the belief that no single class should contain the major portion of the total map error. The map which resulted from this grouping procedure is shown in Figure 19. The BAI value of .91560, although lower than two of those attained earlier, is still very high, indicating that this technique might be productive in grouping other data sets.

Six different modes for manipulating data were attempted in the search for a high BAI value. In addition, BAI indices were calculated in the search for the TAI and OAI values, so that in the total manipulative process eight different modes were attempted.

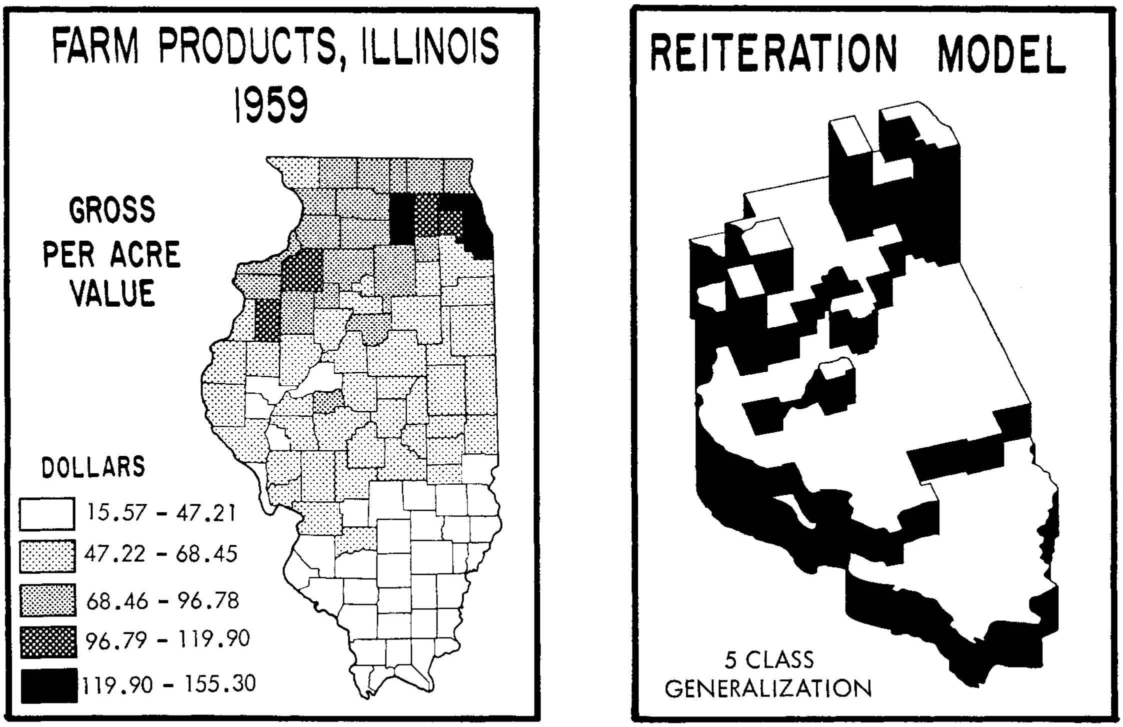


FIG. 20. This is the most accurate five-class map of the Illinois data for the boundary map-user, with a BAI value of .94884.

Prior to, between, and after each of these types of data processing the final classes were reiterated. Thus, nine reiterative cycles were accomplished. A generalization obtained in one of these reiterative cycles surpassed any previously discussed insofar as the boundary map-reader is concerned (Fig. 20 and Table 8). The parentage of this reiterative map is unknown since it could have occurred in any of the reiterative cycles, but it probably was developed in the reiteration which followed the equal average deviation mode. The BAI for this map is .94884.

The Derivation of the Highest MAI

All index values (TAI, OAI, BAI, and MAI) were calculated after each manipulative move. Each of these was compared with the indices obtained for each new set of classes, and the set of classes achieving the highest value for each index was stored. Map accuracy indices were calculated throughout the seventeen cycles of manipulation. At the end of the complete cycling, the classes retained in the MAI storage are the set which achieved the highest composite accuracy index or MAI. The map shown in Figure 15 attained this value (highest MAI).

This generalization will be selected by most map-authors to be presented to their readers, because it is the best *overall* representation of the data set. The data processing described here produces four very accurate generalizations, however, and if the map-author is convinced that his readership will be primarily interested in one type of usage he can select the representation which best suits that need. The MAI values for the grouping procedures discussed are given in Table 8. Three techniques produced MAI values which were higher than those attained by traditional methods (Table 5).

Resume: New Techniques for Generalizing

The generalizing techniques presented in this section are designed to minimize all types of choroplethic map error. Some success is clearly shown by comparisons of the indices in Tables 5 and 8. The OAI, TAI, and BAI values for the most accurate new generalizations exceed those of the traditional processes by approximately three percentage points.

Readers, at this point, may wonder whether the gains in accuracy which have been achieved are of practical value and whether they are applicable to all distributions. Further, they may ask why all of the Illinois maps have been limited to five classes. These topics will be taken up in the following section of the paper, but a synopsis of the generalization procedures and a discussion of their relation to computer processing seems to be appropriate at this time.

The methods of measuring error and the generalizing techniques developed in this paper would be extremely expensive and time-consuming if performed on a calculator one step at a time. Care has been taken, therefore, to organize and develop each part of the procedure so that it can be programmed and accomplished by computers. At this time an integrated computer program to accomplish the total task is unavailable, but a partially completed routine can be used until this objective is reached.[26]

THE LEVEL OF GENERALIZATION

The map is a communicative device and, like all forms of communication, it can be presented at different levels of complexity. In its simplest or most generalized form the Illinois data can be grouped into one class which is represented by the arithmetic mean. Communicators often find a statement such as "On the average, . . ." very useful, but from the cartographic viewpoint this is rather meaningless, as shown by the mean map (Fig. 11). On this map a single shading pattern is used, and the reader is thus unable to perceive surface detail. On the other hand, the data model (Fig. 1) displays the distribution without any generalization whatsoever, and every minor nuance in the configuration of the surface is available to the reader. Whether he is able to perceive all of these surface details is open to question. This would become even more problematical if the ungeneralized data were presented on a one hundred and one class choroplethic map.

Map perception studies indicate that readers are unable to discriminate between patterns when more than ten or eleven are used on a choroplethic representation.[27] Thus, from the practical point of view, map-authors are more or less obliged to present limited generalizations, and the number of classes they select usually ranges from two to ten. Five classes were utilized for the maps presented here, because that number seemed satisfactory for the counting number map. To make the other maps directly comparable with the counting number map, five class levels were maintained throughout the set. Some map authors use as many classes as they can symbolize, whereas others recommend using a very few. The first group disregards the psycho-physiological limitations of the eye and the mind, since their desire for detail supersedes their wish to provide clear boundaries or understandable overviews. Tobler, on the other hand, maintains that more information may be transferred by simple maps with few classes than by complex maps with many classes.[28] Whether Tobler is correct has yet to be proven, but it is clear that map-authors have often set their generalization levels rather subjectively. It is our contention that better choroplethic maps would result if the accuracy of each generalization level was known before the map-author made his decision.

The Map Accuracy Curve

The Illinois data were processed through the reiterative and forcing procedures to create seventeen levels of generalization for the tabular map user. The TAI score for each of these maps is shown in graphic form in Figure 21. The tabular accuracy index increases as the level of generalization decreases with an

[26] This study was facilitated by a generous grant from the Computation Center of the University of Kansas. We are indebted to F. J. Rohlf, J. R. Kispaugh, and R. Bartcher, who developed the NT-SYS computer program which was used to develop our linkage maps. The program described in this paper can be obtained from the Institute of Social and Environmental Studies at the University of Kansas. The program entitled CARTSTAT I was initially developed by Fred C. Caspall and refined and implemented by Carl E. Youngman. It is available from the Institute, which will also quote costs on processing data for other departments or individuals.

[27] R. L. Williams, "Map Symbols; Equal Appearing Intervals for Printed Screens," *Annals*, Association of American Geographers, Vol. 48 (1958), pp. 132–39, and G. F. Jenks and D. S. Knos, "The Use of Shading Patterns in Graded Series," *Annals*, Association of American Geographers, Vol. 51 (1961), pp. 316–34.

[28] Waldo R. Tobler, personal communication.

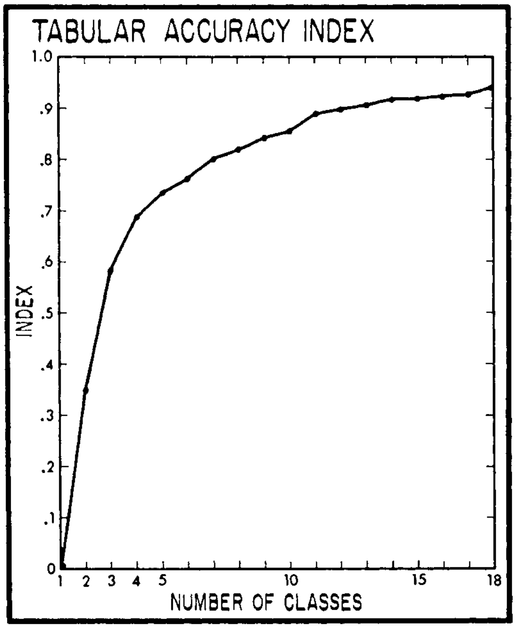


FIG. 21. The decreasing rate of improvement in tabular accuracy is shown by this curve. It is irregular because the data used for these generalizations were a broken series of values (Table 1). Complete or unbroken series of values are rare in mappable data.

increasing number of classes. This increase in accuracy is uneven, however, and the improvement in accuracy increases at a decreasing rate. The unevenness of the curve is to be expected, because the data for Illinois are broken, and these irregularities influence the class groupings. The decreasing rate of improvement is also expected, and similar results would probably be found if other data sets were processed in similar fashion. However, one should not expect to find identical values for other data sets, because each set is more or less unique in the arrangement of intensity values.

Although not shown here, the measures for the other accuracy indices (OAI, BAI, MAI) follow the same trend line pattern as the TAI index (Fig. 21). Unlike the TAI index, however, these other values were obtained by using a variety of generalizing modes. The best two-class MAI for the Illinois data was obtained by using the average deviation generalizing modes, the best three-class MAI by the blanket generalizing mode, the best four-class MAI by the reiteration mode, and the best six-class MAI by the centroid linkage mode.

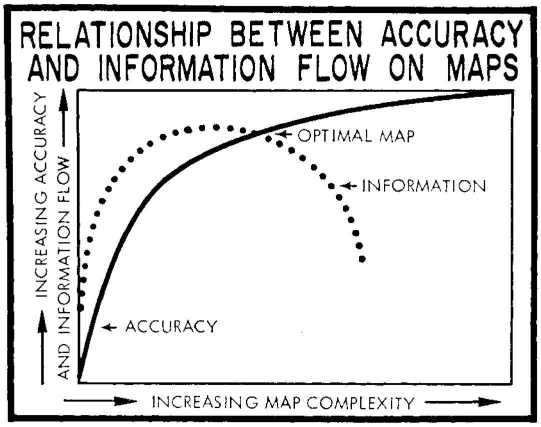


FIG. 22. The accuracy curve has been defined in this paper but the information flow curve (dotted line) is theoretically based. The optimal representation of a distribution will be achieved when both of these map functions are maximized.

Increasing numbers of classes decrease the level of generalization on a choroplethic map and make the map a more accurate representation of the data model. This is only one aspect of map accuracy, however, and as class numbers are increased there is an associated increase in the complexity of the visual patterns which result. This significant aspect of visual static is discussed in the concluding section of this paper.

CONCLUSIONS

In this essay we have been primarily concerned with error in choroplethic map construction. As a result, our focus has been upon the definition, measurement, and reduction of the errors associated with classing data for choroplethic maps. We have not, on the other hand, provided the cartographer with a measure of the information carrying capacity of a map. Neither have we investigated whether readers perceive overviews, nor to what degree their understanding of the distribution may be inhibited by visual static. These are areas for future research, but one is of particular significance, since it sets the framework for an understanding of the choroplethic map as a communicative tool.

An increase in visual static clearly is associated with an increase in the number of classes on choroplethic maps. More detail of areal pattern, boundary, and tabular information is packed into the six-class map than into the two-class map. The problem arises, however, whether the additional information provided by less generalization is perceived and understood by the reader of the more complex map.

Theoretically one can assume that less generalization, i.e. more choroplethic map classes, will first be associated with an increased information flow. It also follows that there is some level at which additional information added to a map is not perceived. Furthermore, a map can become so complex and so riddled by visual static that it no longer functions as a communicative device. In graphic form these concepts trace a U-shaped information flow trend line (dotted line, Fig. 22). The accuracy curve of a series of choroplethic generalizations is a curvilinear trend line (Fig. 21). When this is superposed upon the information curve the two trend lines intersect at a point (Fig. 22). This point locates the level of generalization which will achieve an optimal map, since it is the point at which it is possible to attain both maximal accuracy and information carrying capacity. Such an optimal map is beyond reach for the moment, because we have no measure of the amount of information that is taken from a map by a reader.

A map can be viewed as an information system into which authors input ideas. These ideas are channeled to the reader through both space and time where they are output to a map reader. If the system is efficient input concepts will be output with a minimum of loss and aberration. This paper has been designed to aid map authors in improving inputs into the choroplethic map information system. To be of maximum utility the error reduction system discussed here must be placed in a complete information system. This means that further research on choroplethic map outputs is needed before cartographers can attain optimal representations.