ACCURACY, COMPLEXITY, AND USERS KNOWLEDGE FOR CLASSIFICATION METHOD IN CHOROPLETH MAPS

Behdad Ghazanfari
National Cartographic Centre of Iran (N.C.C.),
P.O.Box 13185-1684, Tehran,Iran
Fax: ++98 21 6001971

Abstract

The classification method in choroplethic mapping should always start with the consideration of the users of the map. If a sophisticated method is selected for a group of people who are not experienced in map reading, the information may not transfer properly, or perhaps a wrong perception of the information may occur. The second important issue in classification is accuracy, which depends on the purpose of the map. Therefore the cartographer should identify the purpose before he or she classifies the data. Complexity of the map pattern is another important factor, because it is at least partly influenced by the classification method and the number of classes.
In an attempt to increase choroplethic map effectiveness, an interactive computer program was produced to integrate the above mentioned factors in selection of classification method.

1 - Introduction

The classification problem in choroplethic mapping is one of the most interesting areas of research in thematic cartography. It is quite a significant and sensitive step in the whole production process of this type of maps. If the map maker chooses a system of classification without analyzing the characteristics of the data and the other aspects which influence the effectiveness of maps, probably the result will not provide the user with the desired information of the spatial data. An inconvenient method of classification may result in a weak or perhaps a wrong impression of the spatial distribution, which the cartographer did not intend to give the map user.

There exists various opinions of many authors on how the selection of class intervals in choroplethic mapping should be accomplished. Different methods of classification lead to various appearances of the maps, and therefore result in different interpretation. Some methods of classification provide more accurate representation of the data than other methods. In some systems the class limits can be determined easily, while in other methods it is complicated and computer programmes are required. Finally, the class limits resulting from some systems are easy to be memorized and matched from the legend to the map or vice versa.

Selection of the number of classes in choroplethic mapping is also a very important factor. However, some authors such as Tobler (1973) and Peterson (1979) believe that data in choroplethic mapping should not be classified because it decreases the accuracy of the map. It is clear that the number of classes greatly affects both the accuracy and readability of the map. It has to do with the question of how quickly and accurately people can identify to which class an areal unit belongs (Gilmartin and Shelton, 1989). The greater the number of classes, the less the accuracy with which map user can match a particular areal shading on the map to the same shading in the legend.

In this paper the classification methods are grouped in some general categories, the traditional data analysis which may assist the cartographer to select an appropriate classification method will be discussed and some fundamental considerations in the selection are explained. The accuracy and pattern characteristics of choropleth maps are briefly reviewed and finally the program which has been written in order to take care of those aspects will be explained.

2 - Data classification methods

Classification is in fact a generalization process. Just as all generalization processes, classification has to be carried out based upon some specific rules and concepts. Every classification method categorizes the data in a different way.

Theoretically, there exist an infinite number of classification systems, but only few of the has been utilized in choroplethic mapping. These methods can be grouped into three categories. The first group includes the classification systems which yield classes with irregular widths. The second group consists of the methods by which constant class width are obtained. And the last group comprises the methods which results in systematically increasing sizes of the classes.

2.1 - Data analysis for classification

The first step in the traditional approach to select a classification method is to put the data in an ascending order, from low to high. By doing so, the data will be prepared for subsequent procedure, besides, it helps to have an idea of the range of the whole data set. After arraying the data in order to visualize the characteristics of the distribution some graphs such as "scatter diagram" may be constructed. A scatter diagram is created to signalize the irregularities in the distribution of the data.

To construct a scatter diagram, a simple scaled line is drown. This scaled line must be long enough to cover the whole range of the data. Each value of the data set is plotted alongside the scaled line in a dot form. The scatter diagram may demonstrate a number of clusters of dots which are separated by empty spaces. These dot-less spaces are called breaks, because gaps of the data values exist in those ranges.

2.2 - Some important consideration in data classification

For classifying quantitative data, traditionally there exist several rules or considerations which should be taken into account by cartographers and most of them were suggested by Jenks and Coulson (1963), as follow:

a) The classification should encompasses the full range of the data,

b) Classes may not overlap,

c) The data that fall into each class should be harmonized. In other word, Classes should be internally homogenous and externally heterogenous,

d) empty classes are allowed to be included in the legend,

e) The accuracy of classification should not exceed the accuracy of the original data,

f) Round-off class limits are better understood and memorized by map users,

g) If possible, a logical relationship exist between the class sizes,

h) Complexity of the map pattern should not disturb the map user.

2.3 - Methods using irregular class widths

Natural break

When looking at the scatter diagram, one may observe distinct breaks in some parts of the data set. In this cases class intervals could be selected in such a way that harmonized clusters of the data are separated. It is ideal that the data with similar characteristics fall into the same class and those with dissimilar are separated.

According to Paslawski (1984), the advantage of using this method is that, it can "...provide instance and visually appealing information about the mapped set". Evans (1977) criticized this method as, it attributes "significance" to very minor troths in the histogram. On the other hand he recommended the use of this method when "...a distribution were demonstrated to be significantly multimopdal...". Smith (1986) in an extensive attempt found that natural break method is particularly unreliable for classing normal distribution.

Quantiles

If it is decided to assign equal numbers of observations to each class, quantiles method can be used. The total frequency of values is divided by the number of classes. Having arranged the data set in an ascending order, the assignment of the observations into the classes is simply done.

Since this method is based on the number of occurrences and not on the magnitudes of the data, it can be very useful for ordered data presentation (Evans 1977, Robinson et.al.1984). By inspecting the scatter diagram one can ensure whether relative similarity inside each class is obtained or not. The main shortcoming of quantiles is that the class limits are irregular between maps showing the same variable for different areas or times.

Equal areas

Instead of obtaining an equal number of occurrences per class, yielding more or less equal areas covered by each class is the concept of equal areas method. To apply this method, information about sizes of the enumeration units must be available.

All of the advantages and drawbacks of the previous method are valid here. When the sizes of the enumeration units differ, which is common in most cases, this method creates a more pleasing image from the user's point of view because the map is covered by almost equal areas for each shading.

Nested-means

The mathematical means divides a numerical array of the data into two classes, and the means of each of these two classes yield four classes with smaller intervals. The result is a series of classes derived from a

nested hierarchy of means. The more closely spaced the values in a data range, the narrower the classes. Scripter (1970) strongly supports this method as, it has the desirable features of being objective and satisfying all requirements needed for class intervals.

The most important advantage of this method is that no empty class can be created, and at any level the distribution of the data can be said to be equilibrium. The latter is due to the characteristic of the mean which is the point of the minimum variance, and therefore the most representative of the data values. The disadvantage or inflexibility of this method is the lack of applicability in situations when the number of classes is not 2^m.

Jenks' optimal

This method was introduced by Jenks in 1977. It incorporates a consistent logic to create classes that are internally homogeneous while assuring heterogeneity among them. For most data sets this method requires availability of computer. After defining the required number of classes, an arbitrary set of class intervals is obtained and the square deviation of each class mean (SDCM) is calculated. An observation on the margin of each class is then removed on an adjacent class. If this transfer reduces the SDCM the observation is left in its new class; if not, it is reverted. This procedure is repeated until the SDCM is minimized.

Coulson (1987) point out that Jenks' optimal method has "...a well defined objective, and a means of objectivity comparing the resulting class intervals for the same data set". However, it should be realized that Jenks' optimal method sometimes ignores the relatively small but interesting differences in data values at the lower end of the range. Besides, in spite of its high accuracy, most of the map users are not likely to be pleased with the irregular class sizes and are not able to understand the concepts behind this method.

2.4 - Methods using constant class widths

Equal steps

The total range of the data is subdivided into the number of classes in such a way, that each class have an equally sized interval. A significant characteristic of this method is the ease of its class interval calculation and also the advantage of easy perception by the user is indisputable. However, when the frequency histogram is not rectangular, use of this method will lead to the creation of some classes containing most of the observations and some which are hardly used.

Standard deviation

Once the standard deviation of a distribution is computed, two class intervals can be determined by adding and subtracting one standard deviation from the mean. Then, every time one standard deviation will be added to and subtracted from the most recently class limits.

Evans (1977) applied this method to a normal distribution using several coefficient s of standard deviations and different number of classes. He found that with any number of classes, when the width of the classes is larger than one standard deviation, few measurements fall into the tail classes. He also found that all classes with the size of 0.5 to 0.6 standard deviation contain almost the same number of observations. He concluded that "...it may be eventually possible to standardize this factor for given number of classes". But this author believes that since in mapping we usually deal with irregular

distribution, we should not base our conclusion on the normal distribution analysis. When the number of classes selected several multiplications of standard deviation should be tested and according to some criteria such as frequency per class the best coefficient should be chosen.

2.5 - Methods using systematically increasing class widths

Arithmetic progression

In this method the widths of the classes increase with a constant value. In fact the aim of using an arithmetic progression is to obtain approximately equal frequency classes, and at the same time obtaining a range of successive classes which can be easily described in mathematical terms.

Geometric progression

When class intervals are calculated in such a way that the upper class limit are always a given number of times larger than the lower class limit, a geometric progression has been used.

The difference between geometric and arithmetic progression methods is, that when the discrepancies between the lower, middle, and upper part of the data range, the former method approximates the distribution better than the latter one.

3 - Accuracy in choroplethic mapping

Many studies have been carried out in order to find out how accuracy of choropleth maps affects their effectiveness. Thijs (1991) puts the burden of finding the wright balance upon the shoulders of the cartographers. It is clear that finding the best method for optimizing error is a difficult task which requires a good knowledge of cartography, analysis and understanding the nature of data.

Among cartographers, Jenks has made a significant contribution in moving choropleth maps from highly subjective works of art towards products of a scientific approach to select class intervals. Jenks and Caspall (1971) classified the concept of accuracy to "overview", "tabular", and "boundary". Overview accuracy becomes significant when the map user seeks an overview of the statistical distribution from a choropleth map and has a volumetric dimension. When the map users may think of maps as areal tables which are used as a source of specific information about a place, tabular accuracy becomes important. It has an areal dimension. If focus upon the boundary lines of the enumeration units are made and comparison of the boundaries with those that are held as the mental images of those which occur on other maps, boundary accuracy play the important role. It has a distance dimension.

The "composite accuracy" index which takes the three errors into account were defined by Jenks and Caspall. This index is comparable whit calculation of the length of a vector in a three dimensional coordinate system. The axes of the coordinate system correspond to the overview, tabular, and boundary accuracy. This author believes that the composite accuracy is not useful due to the fact that the three measures of accuracy are not commensurable. Each of them represents different concept of errors and has different dimension. Hence combining them in one formula is meaningless.

3.1 - Homogeneity accuracy

In order to achieve a banket of error which is uniformly distributed over the entire surface of the map, the

data should be grouped in such a manner that the average deviation of the classes is equal or as equal as it is possible. More recently Jenks devised a measure for evaluating the level of optimization. It is called "Goodness of Variance Fit" (GVF). This index is calculated as follows:

$$SDAM = \sum (X_k - M)^2 \qquad SDCM = \sum \sum (X_{kj} - m_j)^2$$

$$GVF = 1 - (SDCM / SDAM)$$

where SDAM is the standard deviation of the array mean, SDCM is the standard deviation of class mean, X_k is an individual data independent from any class, M is the mean of the total observations, X_{kj} is an individual observation in class j, m_j is the mean of n_j observations in class j.

The goodness of variance fit (GVF) can vary from 0.0 to 1.0. the goodness of variance fit can be applied to any set of class intervals an d comparisons can be made among different classification methods with the objective of optimal accuracy. Coulson (1987) in his experiment on GVF discusses that the goodness of variance fit is the most powerful test yet advanced for comparing and determining the most accurate map with a given number of classes.

4 - Complexity measurements

Obviously, to fulfil the purpose of the map, the final appearance of the choropleth map is always a very important criterion considered by the cartographers. Complexity appears to be the result of four somewhat independent factors: 1) shape, size, and number of ares on the map; 2)spatial variation of the data; 3) number of classes; 4) classification method. The assumption that subjective visual analysis of spatial association is not sufficient as a basis for the support of theoretical constructs, has been the reason for developing the quantitative measures by several researchers.

Muller (1976) suggested three ratios for calculating the visual complexity of choropleth maps by using graph theory. A graph is defined as a collection of "faces", "edges", and "vertices".In a choropleth map those correspond respectively to enumeration units, boundaries between units, and joining points between units. The Muller's complexity indices after applying classification were: 1) the number of faces over the total possible faces; 2) the number of edges over the total number of edges; 3) the number of vertices over the total number of vertices. He also suggested a modification of the edge index in which the length of the edges were substituted for the number of edges. By applying this modification, the differences in size of the enumeration units could partially be taken into account. Lavin (1979) and MacEachren (1982) in their studies found that there exist a very high correlation among the three Muller's measures of complexity. Therefore, only the modified edge ratio could be an appropriate index for map complexity because it can compensate the large variations of face size.

5 - Description of the classification program

A computer program for the selection of classification method was written. In this program three

important issues in choosing a classification method was integrated. The first issue is the group of users and map use. The second issue is accuracy of choropleth maps and the last one is the complexity of map. The program is an interactive one. It assumes that the data have been stored in two files. The name of the main data, the areal data, the level of knowledge of the map users, and number of classes are the questions which have to be answered by the user of the program. If the level of knowledge of map users are specified as "elementary", all classification methods leading to creation of irregular class widths will be ignored. If the level Is introduced as "intermediate", only Jenks' method will be skipped and for "advanced" level all methods will be taken into account. Class interval computation will follow and for each method goodness of variance fit is computed. Map complexity measurements will be carried out in ILWIS software environment for every method. two ranking of classification methods are made: according to accuracy from high to low and according to complexity from low to high. Actually, the second arrangement is according to "simplicity". The method which is the most accurate one will be selected provided that its simplicity rank does not deviate from its accuracy rank more than +3. This value can be changed interactively. If more weight is intended to be given to simplicity (complexity becomes more important), the value +3 can be changed to +2, +1 or 0, and if less weight is decided to be given to simplicity the value can be increased.

Another application of the program could be, to find the optimum number of classes for a particular classification method. Several maps with different number of classes could be produced by ILWIS using an specific classification method. The complexity indices of every map may be computed and ranked. These indices can then be evaluated, taking into account that a high number of classes results in a more accurate map. The final decision on the number of classes will depend on the rate of increase in the complexity measures.

6 - Conclusion

For selecting an appropriate method three important criteria have to be taken into account: map user's knowledge of map reading, accuracy of classification method, and complexity of the map pattern. In a program these criterion were integrated.

Map users were categorized into three levels according to their knowledge: *elementary, intermediate,* and *advanced.* It was suggested that methods leading to irregular class widths should not be selected for the elementary level of user. At the intermediate level Jenks' optimal classification should not be used. For advanced level all methods can be utilized.

Although overview, tabular, and boundary accuracies might be more useful in specific situations, *Goodness of Variance Fit* was used as the measure for accuracy, because often the use of the map is not known in advance.

Complexity is the result of four factors: classification method, number of classes, geographical parameters of the areal units, and the distribution of the data. The cartographer can only manipulate complexity by changing the classification method and / or the number of classes. In this study *Modified edge ratio* was applied to measure the complexity of maps.

7 - References

Coulson, Michael R.C. (1987) "In the matter of class intervals for choropleth maps: With particular reference to the work of George F.Jenks", *Cartographica, Vol. 24, No.2, pp. 16-39*

Evans, Ian S. (1977) "The selection of class intervals",
Transactions, Institute of British Geographers (new series), Vol. 2, No. 1, pp. 98-124

Ghazanfari, Behdad (1993) "Selection of the most appropriate classification methods in choroplethic mapping",
Unpublished M.Sc. Dissertation, ITC, The netherlands, Enschede

Gilmartin,P. and Shelton E. (1989) "Choropleth maps on high resolution CRTs / The effect of number of classes and hue on communication", *Cartographica, Vol. 26, No.2, pp. 40-52*

Jenks, George F. (1967) "The data model concept in statistical mapping",
International Yearbook of Cartography, Vol.7, pp. 182- 188

----------and Caspall,F.C. (1971) "Errors on choroplethic maps: definition, measurement, reduction",
Annals of the Association of American Geographers, Vol. 61, No. 2, pp. 217-244

----------(1977) "Optimal data classification for choropleth maps",
Occasional paper 2, Department of Geography, The University of Kansas, Lawrence, Kansas

Lavin, S. (1979) "Measures of pattern complexity for choropleth maps",
Unpublished Ph.D. Dissertation, The University of Kansas, Lawrence, Kansas

MacEachren, Alan M. (1982) "Map complexity: comparison and measurement",
The American Cartographer, Vol. 9, No. 1, pp.31-46

Muller, Jean-Claud (1976) "Number of classes and choropleth pattern characteristics",
The American cartographer, Vol. 3, No. 2, pp.169- 175

Paslawski, Jacek (1984) "In search of a general idea of class selection for choropleth maps",
International Yearbook of Cartography, Vol. 24, pp. 159-169

Peterson, Michael P. (1979) " An evaluation of unclassed crossed-line choropleth mapping",
The American Cartographer, Vol. 6, No. 1, pp. 21-37

Scripter, Morton W. (1970) " Nested-means map classes for statistical maps",
Annals of Association of the American Geographers, Vol. 60, pp. 385-393

Thijs, Greet (1991) "accuracy against simplicity in choropleth mapping
International Cartographic Association (ICA), Proceedings, Vol. 2, pp. 877- 881

Tobler, W.R. (1973) "Choropleth maps without class intervals
Geographical Analysis, Vol. 5, pp. 262-265