PROPOSITIONAL SOLUTION FOR DATA CLASSIFICATION IN THEMATIC MAPPING

Le Minh Vinh

Department of Geography, University of Social Sciences and Humanities,

10-12 Dinh Tien Hoang, District 1, HCMC

E-mail: leminhvinh@hcm.vnn.vn

ABSTRACT

Data classification is one of common problems that mapmakers confront with when they create some kinds of thematic maps such as cloropleth maps, proportional symbol maps, etc... Traditionally, this problem was solved subjectively, based on mapmakers' experiences.

By using Cluster Analysis theory, this paper concerns about the solution for issues relevant to data classification, including determining the number of classes, choosing methods of classification and assessing quantitatively the results.

1. INTRODUCTION

As a consequence of technological changes in Cartography, nowadays every people can get an opportunity to make thematic maps by themselves owing to the support of GIS software packages. This is highly desirable, because of the significant profit people can get from using maps and (therefore) the obvious increase of amount of map-users in their professional works. In the other hand, it is also problematic because there is no guarantee that the resulting maps will be well designed and accurate.

Data classification is one of common problems that mapmakers confront with in creating some kinds of thematic maps such as cloropleth maps, proportional symbol maps, etc.. Traditionally, this problem used to be solved subjectively based on mapmakers' experiences.

Data classification can be described as grouping data based on one or more characteristics by which data in same group will be identified one with others (equally).

That process can be considered as a kind of generalization, which reduces the amount of details in a map in a meaningful way so that it makes the map become more readable and suitable.

Apart from cases when mapmakers try to present something special for their own purpose, generally, it is the "best" choice for data classification when placing like values in the same class (and unlike values in the different ones). Besides these criteria, in socio-economic thematic mapping, it is desirable to concern some specific values such as mean percentage values, norms, threshold values that separate two levels… so that the result of classification will be more meaningful and the reality of map-products will be improved. That work may be rather complicated, but its necessaries and benefits require some efforts…

An attempt was made on setting up as detailed as possible instruction for data classification so that inexperienced mapmakers may avoid unexpected mistakes and create qualified maps for their own works.

2. PROPOSITIONAL SOLUTION FOR DATA CLASSIFICATION

Using Cluster analysis and Statistical theory, an objective approach for data classification was suggested; it consists of issues relevant to data classification, including determining the number of classes, choosing classification methods and quantitatively assessing the results.

2.1. Determining the number of classes

Theoretically, a set of data can be divided into arbitrary number of classes. Naturally, the more number of groups, the better result we will get. However, to make sense, the number of classes should be limited (due to human being’s ability in distinguishing, for easy-to-get requires). Research has revealed that humans can handle up to maximum of 7 or 9 classes to get an overview and understanding a map at a single glance (Kraak M.J, 1995, p. 141).

In determining an appropriate number of classes, the spatial context of the data should be considered. For this purpose, besides supportive visualizers such as point graphs, dispersion graphs or histograms, we can do some preliminary data exploring by:

- **Calculating similarity between data**

  Many different coefficients have been proposed for similarity measurement. (Clarke K. L., 2001). Bray-Curtis coefficient is one of the suitable that can be used in our thematic mapping

$$S_{jk} = 100\left\{1-\frac{\sum_{i=1}^{p}\left|y_{ij}-y_{ik}\right|}{\sum_{i=1}^{p}\left(y_{ij}+y_{ik}\right)}\right\} = 100\frac{\sum_{i=1}^{p}2\min(y_{ij},y_{ik})}{\sum_{i=1}^{p}\left(y_{ij}+y_{ik}\right)}$$

(S_{jk} is similarity index between i^th and j^th objects in data set)

For univariable data, p=1:

$$S_{jk} = 100\left\{1-\frac{\left|y_{j}-y_{k}\right|}{\left(y_{j}+y_{k}\right)}\right\} = 100\frac{2\min(y_{j},y_{k})}{\left(y_{j}+y_{k}\right)}$$

- **Creating similarity matrix that shows similarity between any pair of data**

$$\begin{array}{lcccccc}
 & \mathbf{X_1} & \mathbf{X_2} & \mathbf{......} & \mathbf{X_i...} & \mathbf{X_j.....} & \mathbf{X_n} \\
\mathbf{X_1} & \mathit{1} & & & & & \\
\mathbf{X2} & S_{12} & \mathit{1} & & & & \\
: & & & & & & \\
\mathbf{Xi} & S_{1i} & S_{2i} & & \mathit{1} & & \\
\mathbf{:} & & & & & & \\
\mathbf{Xj} & S_{1j} & S_{2i} & & S_{ij} & \mathit{1} & \\
: & & & & & & \\
\mathbf{Xn} & S_{1n} & S_{2i} & & S_{in} & S_{jn} & \mathbf{1}
\end{array}$$

- **Building dendogram -** a tree diagram illustrates the arrangement of the cluster, produced by cluster analysis. It shows the genetic relationships among the members of a population (set of data).

  In our example, the x-axis represents the full set of data and the y-axis defines similarity level at which two objects or groups are considered to have fused.

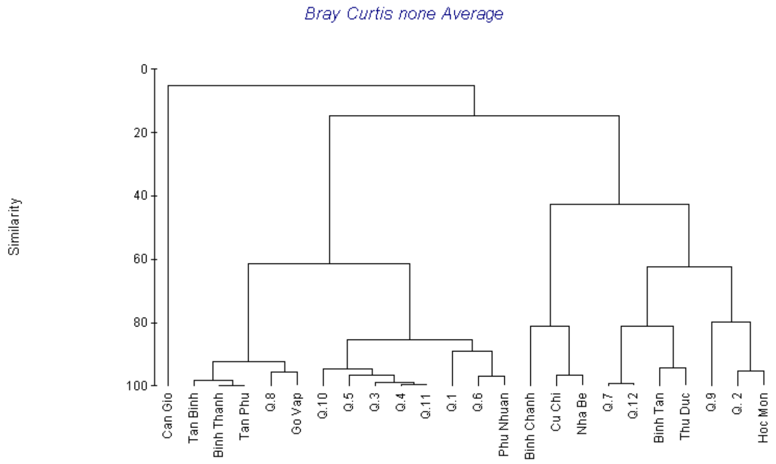

Figure 1. Dendogram shows the cluster analysis result of population density of 24 districts in HCMC using Bray-Curtis similarity coefficient, average linkage

- **Analyzing diagram for choosing the appropriate number of classes**

  The dendogram visually shows that the more classes data set being divided, the higher similarity level we will get. As the number of classes is usually limited from 4 to 9, we should balance while choosing the “second-best” number.

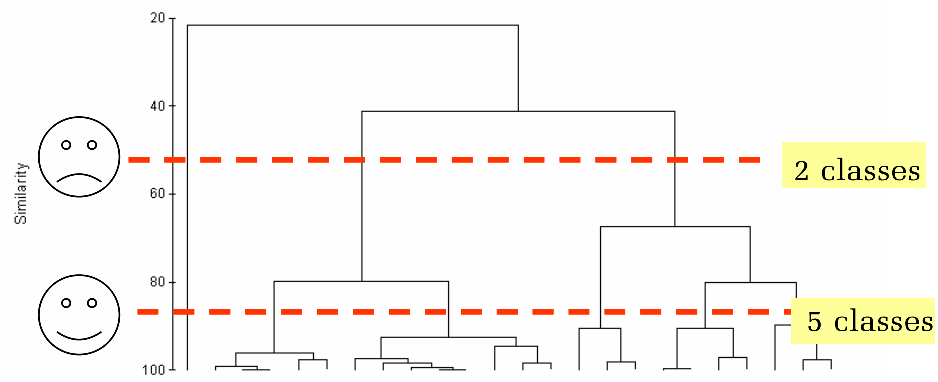

Figure 2. Analyzing dendograph to choosing number of classes

In this example, 5 is the most appropriate number of classes, which will get the similarity up to 85%. Increasing number of classes in this case is not necessary, as it will raise the similarity insignificantly.

2.2. Choosing appropriate methods of classification

There are many different methods of data classification have been used in thematic mapping. Some of them are very simple and easy to use but do not consider how data are distributed along the number line (such as *equal intervals, equal numbers*…). The others are more complicated but do consider spatial context of the data (such as *Standard deviation, Natural Break*…). However, none of them pays attention to above-mentioned (in Introduction) special values, which are worth to concern to get a meaningful result.

A "second-best" data classification method has been suggested that ensures the reflection of both "special values" and raw data (objects in same class are as similar as possible and objects in two classes are as different as possible).

Following this approach, special values should be identified in advance, *the meaningful thresholds would be the edge of classes, the objects in classes would be concentrated around representative values* (if any). After that, an appropriate set of groups will be defined by *minimizing within-groups sum of squares and maximizing between-groups sum of squares*. The whole approach is illustrated by a macro written in Visual Basic so that we can put it to software package as a supportive option.

2.3. Assessing quantitatively the results of classification

Formerly, the results of classification were accepted just by "feeling". Nowadays, as computers can do complex computing tasks, more objective methods of evaluation should be involved.

Besides qualitative criteria such as *ease of computation, ease of understanding*… some quantitative criteria are considered:

- **Correlation coefficient**

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x}) * (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2 * \sum_{i=1}^{N}(y_i - \overline{y})^2}}$$

Where:

$x_1, x_2, x_3..., x_N$: raw data

$y_1, y_2, y_3......, y_N$: classified data

Theoretically, $-1 =< r_{xy} =< 1$, and the closer r_{xy} to 1 the better result of classification we have. However, this coefficient presents only the correlation (similarity) between two sets of data but cannot reflect completely the level of suitability of classification result. Correlation coefficient, so that, usually be used as a secondary criterion for assessing.

- **Goodness of absolute deviation fit (GADF) or Goodness of variance fit (GVF)**

These criteria concern about the distribution of data in groups. They look at both within-group and between-group variance (Terry Slocum, 1999, p. 73)

Using Statistical Theory, these criteria can be calculated by formulas:

$$\text{GVF} = 1- \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x_i})^2}{\sum_{i=1}^{k}(\overline{x_i}-\overline{X})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x_i})^2} = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x_i})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{X})^2}$$

$$\text{GADF} = 1-\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left|x_{ij}-\overline{x_i}\right|}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left|x_{ij}-\overline{x}\ \right|}$$

Where:

$x_1, x_2, x_3..., x_N$: raw data

k: the number of classes

$x_{j1}, x_{j2}, x_{j3}, ..., x_{jn_j}$: data of j^th class and $\overline{x_j}$ is their mean

- GADF (GVF) = 1 when data are not classified (stay as raw data)
- GADF (GVF) = 0 when all data are grouped into only one (group)
- The closer to 1 GADF (GVF) is, the better result we've received

3. CONCLUSION

In spite of its computational complexity, this instruction does give an objective approach for data classification. By putting it into software package as a supportive option, it will be useful for inexperienced mapmakers in creating qualified thematic maps. In a near future, there should be the development of expert system in which the computer can automatically make decisions by using a knowledge base provided by experienced cartographers. This instruction may be considered to add to such knowledge base.

4. REFERENCES

Clarke K. R., Warwick R. M., 2001, "*Change in marine communities – an approach to statistical Analysis and Interpretation*" Primer-E Ltd,

Kraak M. J., Ormeling F.J., 1995, "*Cartography- visualization of spatial data*", Longman Press.

Pham Dinh Van, 2004, " *Textbook in Statistic Theory"*, Construction Publisher, HaNoi.

Terry A. Slocum, 1999, "*Thematic Cartography and visualization*", Prentice Hall.

Vo Van Huy, Vo Thi Lan, Hoang Trong, 1997, "*Using SPSS for Windows in data processing and data analysis*", Science & Technology, HaNoi..

Website: - http://www.clustan.com/what_is_cluster_analysis.html (8-2004, 10-2004)

- http://www.chass.ncsu.edu/garson/pa765/cluster.htm (12-2004)

- http://www.cs.hmc.edu/~fleck/computer-vision-handbook/statistics.html (10-2004)