Perception and Cognition in Univariate and Multivariate Choropleth Maps of Epidemiological Indicators

A Indrayan
Division of Biostatistics and Medical Informatics, Delhi University College of Medical Sciences
Dilshad Garden
Delhi 110095, India
indrayan@ucms.ernet.in

1. Introduction

The major task of choroplethic cartography is to develop methods of mapping that maximise the accuracy of perception and cognition of the spatial distribution of data. An important consideration for this depiction is the choice of data intervals including the number of classes.

Though shades of gray proportionate to individual data values are sometimes advocated (Tobbler 1973) but the common practice is to draw classed maps. This removes some of the cognitive burden of grouping from the reader. But the number of categories should be such that an average reader can reasonably discriminate and, at the same time, the categories should adequately represent the variation in the data. Besides the number of categories, the second problem is the choice of actual cutoffs to divide the data range into classes. Researchers seem to have overwhelming fancy for the "nice" cutoffs such as multiples of 5 or 10. The classes are often chosen equal, though there are examples of classes based on quantiles, on mean-SD and on statistical significance. None of these take care of numerical similarity in the data and two very similar units can belong to different categories, while markedly different can belong to the same category.

We explore various methods of cluster analysis in the context of epidemiological indicators that may overcome the above mentioned problems. These are applicable to the multivariate data as well as and solve the problem of classing in such data. These may increase the perceptional and cognitive qualities of thematic maps.

2. Methods

Heirarchical cluster analysis is a data exploration technique which seeks to divide the data points such that units similar in some sense form one cluster while the dissimilar ones go into a separate cluster. "Natural groups" are discovered that fit the observations.

Different shades in different classes is an essential features of thematic maps. It is naturally presumed by the reader that the units with same shades are similar. Thus the concept of similarity and dissimilarity seems naturally built into the cognitive process of reading a classed map. Thus the classing obtained by clustering of data seems to very well meet the objectives of thematic mapping.

Different methods of cluster analysis perform well on different kind of data. Most comparison studies seem to conclude that average linkage and Ward's methods are better on many data sets than simple linkage, complete linkage, centroid and median method. (see, e.g., Milligan 1980). The methods with least bias are those based on nonparametric density estimation (SAS 1989). If nothing is really known about the structure of the data, as is likely with most epidemiological indicators, this method may be particularly relevant. Details of two such methods — two-stage density linkage and modeclus — are in a SAS Technical Report (1993).

We examined the performance of average linkage, Ward's, two-stage density linkage and modeclus on a variety of epidemiological data sets. Different methods give different results but a consensus among them, when reached, seems to provide stable clusters. We present results for one such epidemiological data set in this communication. This contains four rates of childhood mortality, namely, still birth rate, mortality within a week, mortality from one week to four weeks of life and post-neonatal mortality for States of India in the year 1996 (SRS 1996).

3. Results

The clusters obtained by different methods are shown in Figure 1. The consensus is also shown. Because of multivariate data, graphical method (Indrayan and Kumar 1996) seems to be the only way to find consensus. The choropleth map obtained is not included here because our software does not allow its inclusion in this note. This however is available on request with the author

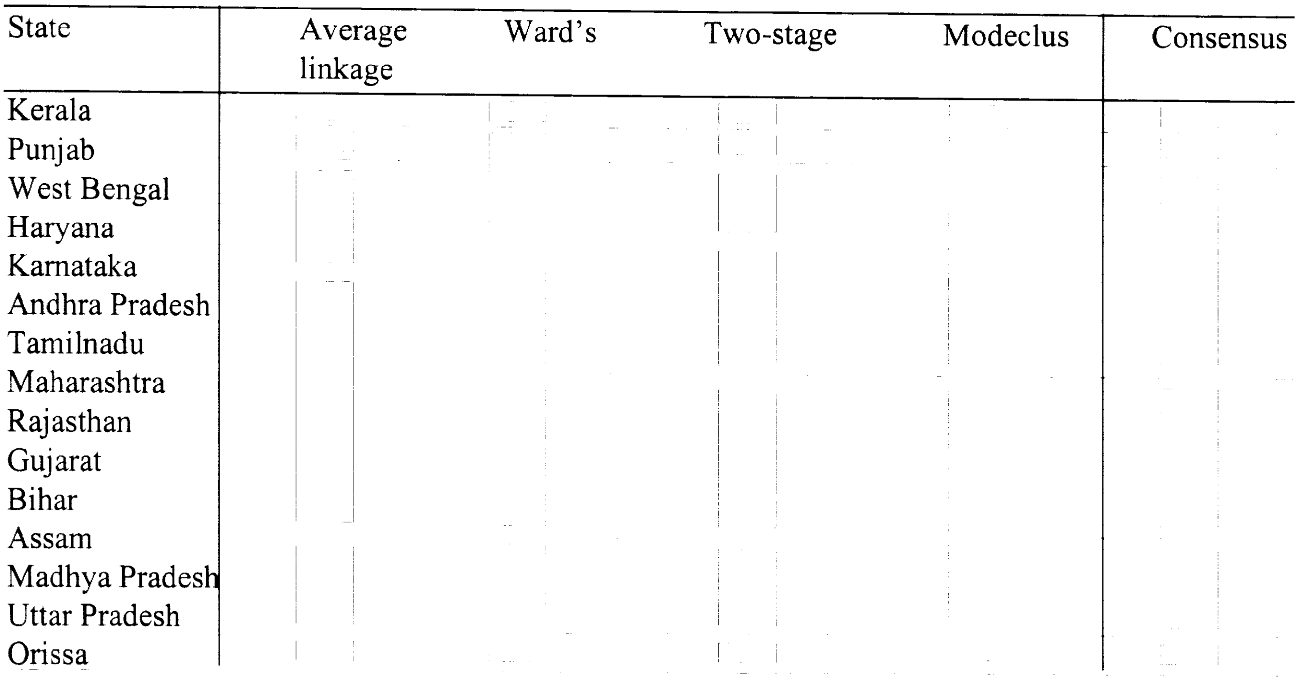

Figure 1. Clusters obtained by different methods on four indicators of childhood mortality - States of India, 1996

4. Conclusions

Our analysis of several data sets, of which one is illustrated above, indicates that classing by consensus among clusters obtained by suitable methods may be preferable in many cases in both univariate and multivariate setups. This seems a definite help in substantially improving the accuracy of perception and cognition from health maps with regard to similar and dissimilar areas.

REFERENCES

Indrayan A. and Kumar R. (1996). Statistical choropleth cartography in epidemiology. International Journal of Epidemiology 25, 181-189.

Milligan G.W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithm. Psychometrika 45, 325-342.

SAS (1989). SAS/STAT User's Guide, Version 6, Fourth edition, V1. SAS Institute Inc., Cary, NC.

SRS (1999). Sample Registration System 1996. Registrar General, India, New Delhi.

SRS Technical Report (1993). SAS Technical Report P-256, SAS/STAT software: The modeculus procedure, Release 6.09. SAS Institute Inc., Cary, NC.

Tobbler W.R. (1973). Choropleth maps without class intervals? Geographical Analysis 5, 262-264.