International Journal of Epidemiology

Vol. 25, No. 1
Printed in Great Britain

Statistical Choropleth Cartography in Epidemiology

A INDRAYAN AND R KUMAR

Indrayan A (Division of Biostatistics and Medical Informatics, Delhi University College of Medical Sciences, Dilshad Garden, Delhi 110 095, India) and Kumar R. Statistical choropleth cartography in epidemiology. *International Journal of Epidemiology* 1996; **25:** 181–189.

Background. The potential of maps in the study of regional variation and similarity in health and in understanding the underlying processes is being increasingly realized. It has thus become important that more care is exercised in drawing health maps and the subjective elements are minimized. Conventional choropleth maps based on quantitative data are mostly arbitrary with regard to the number of categories and the cutoff points. This can lead to substantially different pictures based on the same data set.

Methods. We suggest use of cluster methods to discover 'natural' groups of data points which to a large extent are suggested by the data themselves. These methods can determine not only the cutoff points but also the number of categories required to depict the variability in the data. The methods have natural extension to the multivariate set-up and thus can provide the strategy to construct integrated maps based on the simultaneous consideration of several variables. Since different cluster methods can yield different groupings we propose a simple method to identify cutoffs common to a majority of the methods.

Results. The details of the methods are explained on two real data sets. One is the indicators of mortality before one year of age in India and the other is years of life lost due to premature mortality in different countries. The maps obtained are compared with the conventional maps.

Conclusion. The cutoff points obtained by a majority of cluster methods deserve attention for obtaining natural groups for choroplethic depiction. Maps based on such cutoffs seem to have promise for increasing the accuracy in perception and cognition of regional variation.

Keywords: health indicators, thematic mapping, cluster analysis, natural categories

Maps are powerful tools with which to study the spatial distribution of any phenomenon. They are considered superior to statistical tables in demonstrating regional variation.[1] Traditionally maps are used as an epidemiological tool to identify pockets of concentration of disease. Such maps are available for diseases such as filariasis, leprosy and endemic goitre in India,[2] for gastrointestinal mortality in Argentina[3] and for ischaemic heart disease mortality in New Jersey, USA.[4] Lately maps have also been drawn for health indicators such as the under-5 mortality rates.[5] Maps can be used to forward hypotheses on aetiology, for investigation of these hypotheses, as well as to study the impact of the process and structure of social set-up in determining health outcomes.[6] Their potential in health management is being increasingly realized. It has thus become important that more care is exercised in drawing health maps and that subjective elements are minimized.

The major task of cartography is to develop methods of mapping that maximize the accuracy of perception and cognition.[7] This accuracy depends on factors such as colours or shades used in the map,[8] the geographical unit of choroplethic depiction[9] and the data intervals chosen.[10] In this communication we concentrate on the choice of data intervals for the mapping of health indicators.

Two kinds of data categories are conventionally used in mapping. The first, based on statistical significance is as used by Weiss and Wagener[11] for the standardized mortality ratio (SMR) of asthma. They divide the State Economic Areas of the US into those with an SMR significantly more than 100 and those whose SMR is not significantly more than 100. In such a dichotomy, the areas with a very high mortality get the same depiction as those with not such a high, but significantly higher, mortality. Thus, this categorization has limited value. The second type of categorization is mostly arbitrary. A UNICEF document[12] gives a map of India dividing the States by infant mortality rate (IMR) into ≤74, 75–99, 100–124 and 125+. The number of categories in this case is four. It can be argued that three or six categories would give a better representation. Indeed, in another study,[13] which incidentally is also

Division of Biostatistics and Medical Informatics, Delhi University College of Medical Sciences, Dilshad Garden, Delhi 110 095, India.

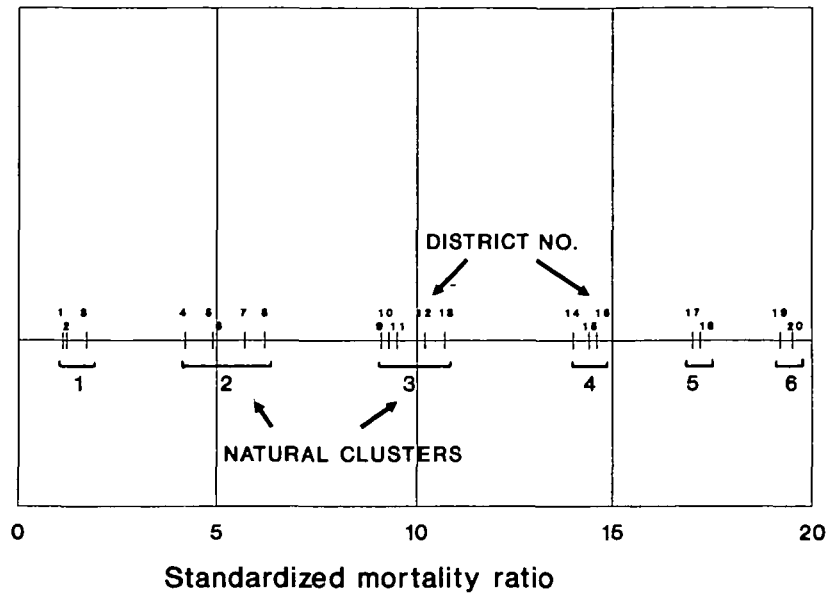


FIGURE 1 *Standardized mortality ratio of breast cancer in 20 districts (hypothetical data)—Conventional cut-points versus natural groups*

supported by UNICEF, the categories used for mapping IMR in India are ≤90, 90–115, 115–140, 140–165 and 165+. The number of categories in this map is five. We have earlier shown that a different number of categories can lead to substantially different pictures.[10,14] Thus, this number needs to be decided on the basis of some objective consideration. There is also the question of deciding the actual cut-points. If we decide to represent the IMR in India in four categories, should they be ≤49, 50–74, 75–99, 100+; ≤64, 65–94, 95–124, 125+; or ≤74, 75–89, 90–104 and 105+? Built into this is the question of deciding on the initial cut-point and of deciding on the width of each category. Conventional maps are arbitrary with regard to these aspects. For example, the World Development Report provides a map of the world showing countries with under-5 mortality in six unequal categories (≤24, 25–49, 50–74, 75–124, 125–174 and 175+). Walter[15] gives a map for Ontario on the incidence of male bladder cancer dividing its counties in quintiles of incidence without explaining why other categorizations such as quartiles are less adequate. Other examples of the use of such arbitrary categories are Pickle and Hermann,[8] Lewandowsky *et al.*[9] and Weiss and Wagener.[11]

We have also shown earlier that different data intervals can lead to different mappings of the same set of data.[10] An additional difficulty with arbitrary categories is that geographic units with very similar rates can undesirably fall into different classes. We illustrate this using hypothetical data on ascendingly ordered SMR for breast cancer in 20 districts (Figure 1). Conventional divisions such as 0–4.9, 5.0–9.9, 10.0–14.9 and 15.0–19.9 lead to four categories. In this categorization, a district with SMR = 1.1 belongs to the same class as a district with SMR = 4.9. But another district with SMR = 5.0 goes into the next group. Common sense dictates that the districts 5 and 6 in Figure 1 with rates 4.9 and 5.0 should belong to the same group. An examination of the SMR reveals that there are six 'natural' groups and not four. The misclassification which occurs through the use of conventional cut-points is clearly visible. Natural cut-points are easily identifiable in this example but in general this will not be so easy. We must use some method to discover natural grouping. Our meaning of natural grouping is that the values tend to be close or similar to one another when they belong to the same group and tend to be apart or perceptionally different when they belong to different groups.

Another problem is in mapping of multivariate data. There have been attempts to depict two or more measurements by using multiple maps,[16] or by using a mixture of patterns and colours.[13] Examples of such maps providing an integrated picture based on

simultaneous consideration of several variables are rare. Some details of these rare examples are given later in the Discussion. Unlike the univariate set-up, inspection of the values in multivariate cases would almost never help in identifying natural cut-points. Walter and Birnie[17] have discussed various other problems in health atlases but they do not mention anything on arbitrary cut-points nor on univariate versus multivariate mapping. They have given guidelines for drawing maps but these too do not take care of the two problems we have highlighted. We propose a mapping strategy based on cluster analysis which takes reasonable care of these deficiencies.

CLUSTER ANALYSIS METHODS

Cluster analysis[18] is a data exploration technique which seeks to divide the data points into clusters such that the units similar in some sense form one cluster while the dissimilar ones go into separate clusters. This procedure is well known as a means of discovering natural groups of data points in the sense that they are suggested, to a large extent, by the data themselves. Thus the number of categories and the cut-points of data intervals are both objectively determined. This method can also be used when simultaneous consideration of several variables in a multivariate set-up is required. However, there are far too many methods of clustering available and there is no unanimity on superiority of one method over the others for any specific kind of data set.[19] These techniques are also notorious for discovering clusters where none really exist.[20] The following is a brief review of the methods and of the relative merits of some of those more commonly used.

Clusters can be overlapping whereby some units can belong to two or more clusters. For example, a patient can be hypertensive as well as diabetic. This corresponds to statistical multiple response. 'Fuzzy'[21] is an example of such clustering. For the purpose of mapping of health indicators, which is the primary focus of this communication, non-overlapping—mutually exclusive and exhaustive—clusters are more appropriate. Non-overlapping methods are of two types—non-hierarchical and hierarchical. The former generates a single partition of the data in order to recover natural groups present in the data. FORGY k-means[19] and MASLOC used by Thielemans *et al.*[22] are examples of non-hierarchical methods. Hierarchical clustering envisages the nested fusion of units of clusters, called agglomerative, or nested fission, called divisive. The CLUE method[22] is an example of the divisive algorithm. Agglomerative methods are generally preferred because of simplicity in understanding. A large number of agglomerative methods are available but the following seven are described[19] as the commonly referenced methods—single-link, complete-link, group average, weighted average, centroid, median and Ward's. Details of these methods are available in Anderberg[18] and Jain and Dubes.[19] The methods differ with respect to the measure of distance between clusters. For example, single-link considers the distance between the two closest units belonging to different clusters as the measure of distance between clusters, while complete-link uses the distance between the two most remote units. As already mentioned, there is no agreement on the superiority of one method over the others. Few theoretical guidelines are available for comparative analysis and no methodology has found wide acceptance. Nevertheless, a large number of comparative studies have been carried out. Golden and Meehl[23] compared the group average, complete-link, Ward's, single-link, centroid and median methods on a particular data set and found the first three outperformed the latter three. In our own simulations,[24] we found complete-link best in not imposing artificial structure on random data. Milligan and Schilling[25] found Ward's method performed best for clusters of equal size, and group average best for clusters of unequal size among complete-link, group average and Ward's methods. Mezzich and Solomon[26] compared 18 clustering methods on four real data sets and found complete-link generally the best. Bayne *et al.*[27] found Ward's and complete-link preferable to the median, group average and centroid methods in a Monte Carlo experiment. Kuiper and Fisher[28] found Ward's method best in their simulations. Thus, the consensus seems to be in favour of complete-link, Ward's and group average methods. Reports of other methods being superior are relatively infrequent. Our intention is not to pass judgement on the superiority of some methods over others and we agree with Jain and Dubes[19] that comparison is a continuing problem for research, requiring more investigations using a range of data types. The above review does, however, help us to decide to limit our investigations to complete-link, Ward's and group average methods only. The following is a brief description of these methods.

The hierarchical agglomerative methods, including the three mentioned above, start with as many clusters as the number of units and do successive fusion till all the units merge into one cluster. These methods do not allow units to separate from clusters to which they have been once assigned. Those two units are merged first which are least dissimilar. Although many measures of similarity and dissimilarity are available, the one most commonly used is the square of the Euclidean distance

$d_{ij} = \sum_p (x_{ip} - x_{ip})^2$ where x_{ip} and x_{jp} ($p = 1, 2, \ldots, P$) are P-dimensional vectors. In the case of one dimension, this reduces to the simple $(x_i - x_j)^2$. To measure dissimilarity between groups of points obtained after such fusion, the metric used by different methods is as follows:

$$\text{Complete link} : D_{hk} = \max_{x_i \in c_h} \max_{x_j \in c_k} d_{ij},$$

$$\text{Ward's method} : D_{hk} = \sum_p (x_{hp} - x_{kp})^2 / (1/N_h + 1/N_k),$$

$$\text{Group average} : D_{hk} = \sum_{i \in c_h} \sum_{j \in c_k} d_{ij}/(N_h N_k),$$

where it is assumed that the i-th unit is in group C_h and j-th unit in group C_k, and that group C_h contains N_h units and C_k contains N_k units. In simple words, the dissimilarity between the two groups of units in the case of complete linkage is equal to the distance between two most remote units belonging to these groups. In the case of Ward's method those two groups are merged which result in least increase in within sum-of-squares. The group average method considers the average of the distance between the units in different groups as the measure of dissimilarity between those groups.

There is some debate on the criteria to decide the optimum number of clusters. The optimum is obviously the stage when the clusters are compact within but distinct from one another. The SAS package[29] gives a large number of criteria which can be used to assess this. We use a simple criterion—the distance between the most dissimilar units of the two clusters being merged at each stage. This is readily computed by the SPSS package,[30] which is the package available to us, and which looks adequate to illustrate our method of determining data intervals for choropleth mapping. A sudden rise in the value of this criterion at any stage relative to the previous stage would indicate that clusters containing very dissimilar units are being merged. This is the stage to stop the agglomerative process and thus obtain the 'natural' clusters.

METHODOLOGY

Our methodology is to obtain clusters in the data set by each of the three clustering methods, plot bars corresponding to each cluster, look for cutoff points revealed by at least two methods and use these cutoffs for mapping health data. This is based on the premise that an agreement on cutoff points between two or more methods out of three is more of an indication that a 'natural' cutoff exists compared to the cut-point revealed by one method alone. This does not imply that we expect to derive benefits from all the methods. The strategy is fairly general and can be adopted even if clustering methods other than those used by us are preferred. Further, the number of methods for identifying clusters obtained by the majority could be five or seven instead of the three used by us. An odd number is required to break a tie when it occurs. The distance measure we use is the square of the Euclidean distance on standardized values but any other measure such as an angular separation or the Canberra metric[31] can also be used. Also, the criterion to decide the optimum number of clusters can be other than the one used by us.

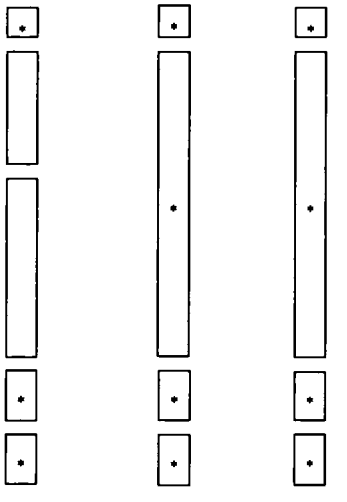
One limitation of the proposed method is that it ignores the geographical contiguity of the areas. For the type of indicators we are investigating such as IMR, the geographical contiguity may not be of much consequence, but for indicators such as incidence of diseases such contiguity could be important. If so, methods of spatial analysis can be used. One review of these methods is by Marshall.[6]

Using the analogy of Mayer,[7] the second limitation of this communication is that we are restricted to the anatomy of geography and not discussing its physiology. We hope that the users of our methodology will find it useful in the more objective appraisal of the patterns and thus be able to suggest more efficient strategies to improve health.

DATA SETS

We illustrate our method on two data sets. The first comprises indicators of mortality before one year of age, namely, stillbirth rate (SBR), perinatal mortality rate (PMR), neonatal mortality rate (NMR) and IMR for the States of India. We consider them in univariate set-up as well as in multivariate set-up. The State is the smallest geographical unit for which these indicators are available in India. These are estimated by the Sample Registration System.[32] These estimates are considered fairly reliable because of a built-in double check system. Our data belong to the year 1990.[32] The rates are generated for 15 major States comprising 97.28% of the population. An advantage with such a small data set is that we can fully illustrate our method. The disadvantage is that State is too big a unit for choropleth mapping. We overcome this in the second data set which is for the years of life lost per 1000 population due to premature mortality in more than 100 countries around the world. These data are for the year 1993 and are given in a recent World Development

TABLE 1 *Clusters obtained by complete-link, Ward's and group average methods on infant mortality rates for major States of India, 1990*



* Similar clusters obtained by at least two methods.

Report[4] of the World Bank. The computation assumes a life expectancy of 80 years for males and 82.5 years for females, and conveys the total burden of mortality in absolute terms.

RESULTS

Let us first consider IMR for States of India. The clusters obtained by complete-link, Ward's and group average methods are shown by bars in Table 1. In this case Ward's method and the group average method incidentally provide the same clusters—thus instantly giving the clusters obtained by a majority of the methods. This agreement may not always be as easily obtained. We illustrate this difficulty later when we consider the multivariate set-up. The clusters obtained by a majority of the methods for all the four indicators, namely, SBR, PMR, NMR and IMR along with their actual values are shown in Table 2. Since the values of these indicators across the States do not follow a similar pattern it is not possible to draw bars. The cluster to which the State belongs is shown by the numbers in the respective columns. Note first that the number of clusters into which the 15 States are divided with respect to SBR is only three, while this number is four for the other indicators. Second, there is no uniformity in the cluster to which a State belongs across the rates. Thus, there is no easy method of depicting all the four rates in one map. The combined situation obtained by simultaneous consideration of all the four indicators is easily obtained by clustering considering all the indicators of equal importance on a standardized scale (mean 0, variance 1). These clusters are shown by bars in Table 3. Different methods yield a different number of clusters and different cut-points. However, at least two of the three methods yield some similar clusters which are indicated by an asterisk (*) sign in the bars. Thus, there is some

TABLE 2 *Clusters of States of India based on stillbirth rate (SBR), perinatal mortality rate (PMR), neonatal mortality rate (NMR) and infant mortality rate (IMR) in univariate set-up by majority agreement on complete-link, Ward's and group average methods*

State	SBR		PMR		NMR		IMR	
	Rate	Cluster	Rate	Cluster	Rate	Cluster	Rate	Cluster
Andhra Pradesh	13.2	1	52.3	2	48.3	3	70	2
Assam	11.4	1	40.9	2	48.1	3	76	2
Bihar	8.6	1	42.8	2	48.9	3	75	2
Gujarat	6.5	1	42.8	2	50.0	3	72	2
Haryana	19.7	2	42.1	2	38.6	2	69	2
Karnataka	18.0	2	57.6	3	51.1	3	70	2
Kerala	10.3	1	20.0	1	12.6	1	17	1
Madhya Pradesh	10.5	1	59.0	3	71.9	4	111	4
Maharashtra	13.8	1	44.7	2	42.2	2	58	2
Orissa	20.9	2	73.1	4	78.8	4	122	4
Punjab	27.9	3	49.0	2	33.6	2	61	2
Rajasthan	7.7	1	45.0	2	52.3	3	84	3
Tamilnadu	11.1	1	47.1	2	43.9	2	59	2
Uttar Pradesh	8.3	1	51.6	2	65.2	4	99	3
West Bengal	16.8	2	42.1	2	37.4	2	63	2

TABLE 3 *Clusters obtained by complete-link, Ward's and group average methods on indicators of mortality before one year of age—stillbirth rate, perinatal mortality rate, neonatal mortality rate and infant mortality rate—States of India, 1990 (multivariate set-up)*

State	Complete link	Ward's	Group average	Consensus
Kerala	*	*	*	*
Punjab		*	*	*
West Bengal				
Haryana				
Karnataka				
Andhra Pradesh				
Tamilnadu				
Maharashtra				
Rajasthan				
Gujarat				
Bihar				
Assam				
Madhya Pradesh		*	*	*
Uttar Pradesh				
Orissa		*	*	*

* Similar clusters obtained by at least two methods.

agreement on Kerala, Punjab and Orissa as individual clusters, and on Madhya Pradesh and Uttar Pradesh together forming one cluster. But there is no agreement on clustering of the other States. In the absence of this agreement, our strategy is to adopt the safer course of putting them together in one cluster. This completes our clustering algorithm. More often than not, the clusters so obtained can be given rank based on the values of the individual indicators. This will help in giving proper shading to various States. Kerala has least value on three out of the four mortalities—thus stands our best with rank 1. Punjab has high SBR and high PMR but low NMR and low IMR (Table 2). Since it too is a cluster by itself, its rank would either be 2 after Kerala or 3 after all the States from West Bengal to Assam in Table 3. Individual values indicate that the former is more likely than the latter—thus Punjab gets rank 2. The cluster comprising Madhya Pradesh and Uttar Pradesh gets rank 4 and Orissa rank 5. The map thus obtained is as in Figure 2. This was drawn with the help of EpiMap program.

The stem and leaf plot of the values of years of life lost per 1000 population for 109 countries is in Table 4. In Table 5 are the cut-points obtained by the three cluster methods on these data. Because of the univariate set-up, it is possible to specify the cut-points. All three methods luckily reveal three clusters but the cut-points are different. The clusters commonly obtained by at least two methods are marked with an asterisk (*) sign. The EpiMap obtained is Figure 3A. For comparison, a map with arbitrary categories is given in Figure 3B. This map has five categories each of width 25.

DISCUSSION

Differences in cluster analysis based world map on years of life lost per 1000 population in Figure 3A and the arbitrary map in Figure 3B highlight the importance of being objective in choosing cut-points for drawing health maps. Our analysis indicates that the countries should be divided into three groups (years of life lost <55, 59–141 and 188) in place of the arbitrary five; or any other categories. The perception of Middle East countries and of African countries obtained from the former should be closer to reality than the one obtained by the latter.

The grouping of areas on the basis of numerical similarity in multivariate data is easily obtained by cluster analysis methods. Thielemans *et al.*[22] used such methods to map female cancer patterns in 43 districts of Belgium but that they say is out of necessity to depict multivariate data rather than by choice. The only other examples of the use of cluster techniques in health mapping which we could locate after searching the MEDLINE and HEALTH databases are the studies of Verhasselt and Mansourian[33] and of Stanfel.[34] The former used unspecified methods of hierarchical and non-hierarchical clustering on a set of 23 health-related indicators and prepared a cartogram of the world. The latter study was based on 48-component cancer mortality data vectors on States of the US. Stanfel proposed cluster methods based on optimization of an objective function and used the obtained clusters to map the mortality. Both these studies have briefly mentioned the need to study the stability of clusters across the methods. Ours is a concrete proposal on the use of bars as a convenient means of implementing this strategy.

Note that there is apparently no study on the use of cluster methods for mapping univariate data on health. The use of arbitrary categories in univariate cases is even more alarming and seems to have escaped

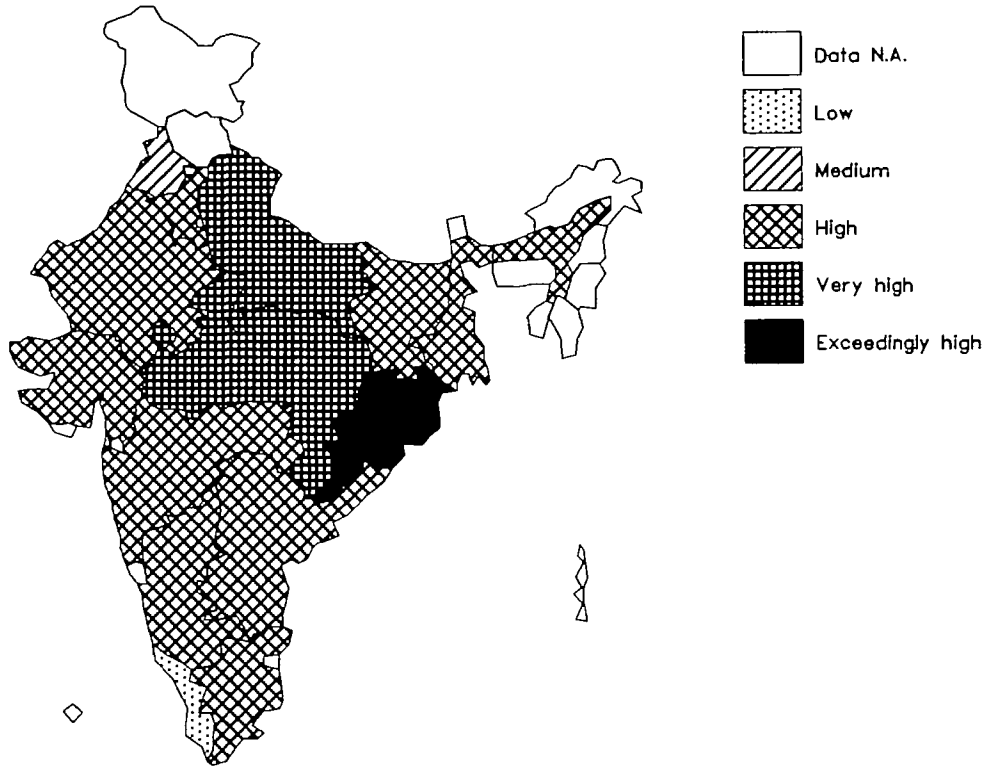


FIGURE 2 *'Natural' clusters of States of India on indicators of mortality before one year of age 1990—stillbirth rate (SBR), perinatal mortality rate (PMR), neonatal mortality rate (NMR) and infant mortality rate (IMR))—multivariate set-up*

TABLE 4 *Years of life lost per 1000 population in 109 countries in the year 1993—stem and leaf plot*

0	789999
1	00000000011111111222223344455556666667789999
2	001122445677789
3	1223677
4	01355
5	059
6	137799
7	499
8	1469
9	389
10	46778
11	024
12	145
13	
14	1
15	
16	
17	
18	8

TABLE 5 *Cut-points on years of life lost (data set 2) by complete-link, Ward's and group average methods*

Complete-link	Ward's	Group average
≤86	≤55*	≤55*
89–141	59–99	59–141
188*	104–188	188*

* Similar clusters obtained by at least two methods.

attention so far. We fill this void and hope that choropleth health maps henceforth will be drawn based on more objective considerations with regard to data intervals. This could enhance the accuracy of perception and cognition which is the basic function of maps. The cut-off points obtained by a majority of the chosen cluster methods deserve attention in determining the appropriate number of categories and their boundaries.

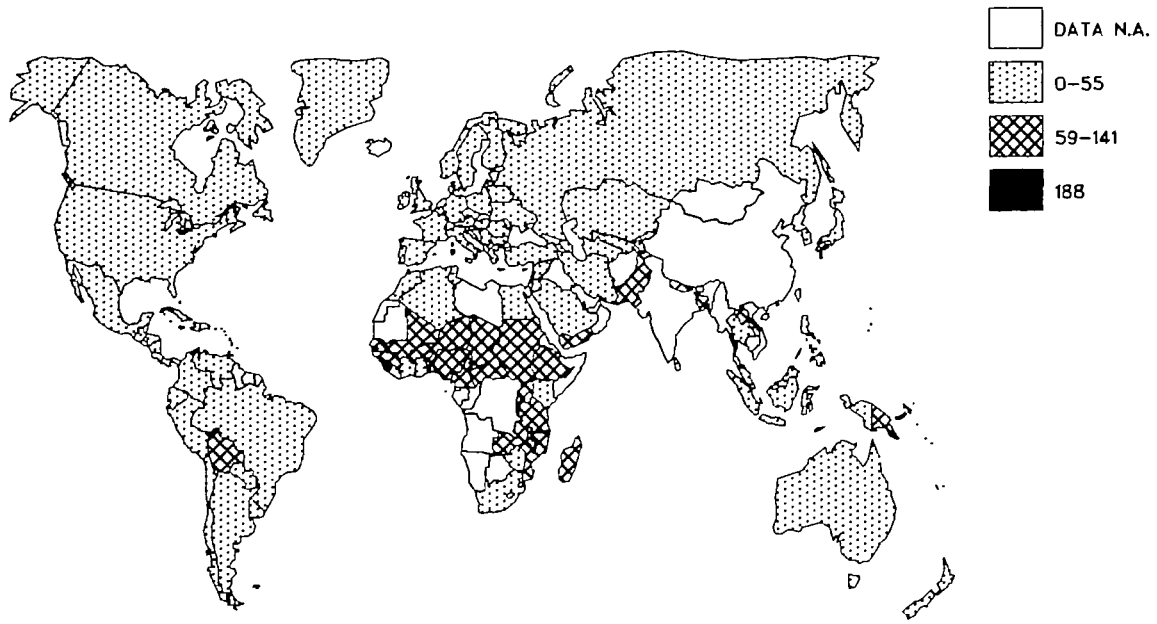


FIGURE 3A *'Natural' clusters obtained by the majority of methods on years of life lost/1000 population in different countries of the world—1993*

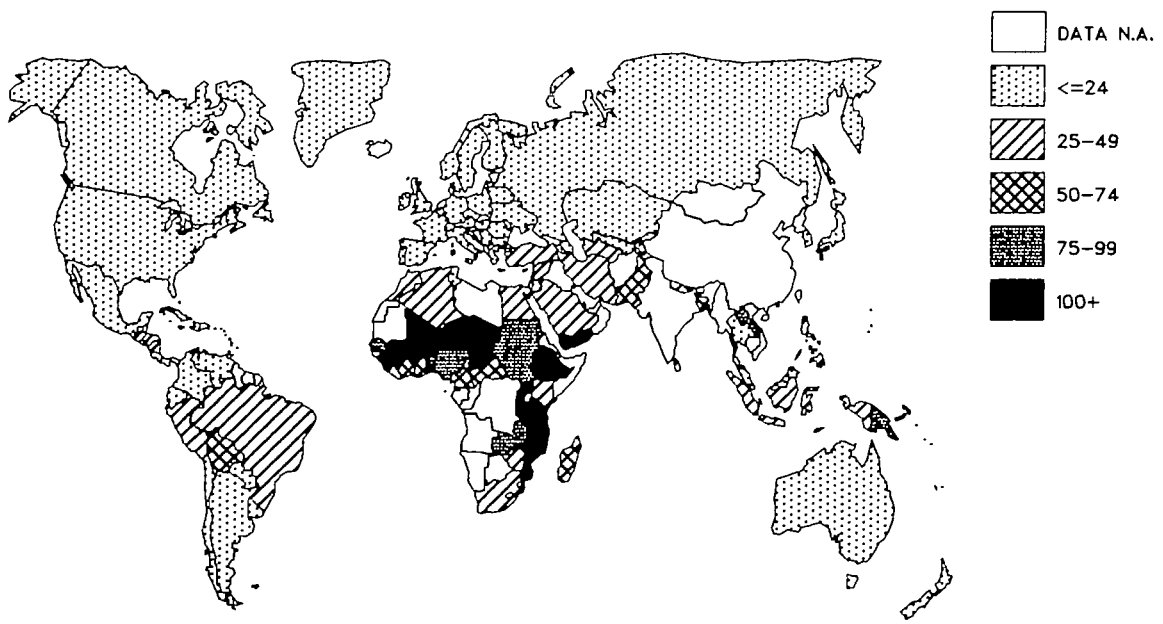


FIGURE 3B *Arbitrary categories on years of life lost/1000 population—1993*

REFERENCES

[1] Vauramo E, Mikkola P, Sippo-Tujunen I *et al*. Co-ordinate based mapping – a new method in health services research. *Med Info Lond* 1992; **17:** 1–9.

[2] Park J E, Park K. *Textbook of Preventive and Social Medicine, 13th edn*. Jabalpur: Banarsi Das Bhanot, 1991.

[3] Matos E L, Parkin D M, Loria D I, Vilensky M. Geographical patterns of cancer mortality in Argentina. *Int J Epidemiol* 1990; **19:** 860–70.

[4] Nagem G R, Hutcheon D, Fuerman M. Changing patterns of ischaemic heart disease mortality in New Jersey 1968–82, and the relationship with urbanization. *Int J Epidemiol* 1990; **19:** 26–31.

[5] World Bank. *World Development Report 1993: Investing in Health*. Oxford: Oxford University Press, 1993.

[6] Marshall R J. A review of methods for the statistical analysis of spatial pattern of disease. *J R Stat Soc A* 1991; **154 (Part 3):** 421–41.

[7] Mayer J D. Medical geography—An emerging discipline. *JAMA* 1984; **251:** 2680–83.

[8] Pickle L W, Hermann D J. The process of reading statistical maps: The effect of colour. *Statistical Computing & Statistical Graphics Newsletter* 1994; **5:** 1–16.

[9] Lewandowsky S, Hermann D J, Behrens J T, Shu-chen Li, Pickle L, Jobe J B. Perception of clusters in statistical maps. *Appl Cognitive Psych* 1993; **7:** 533–51.

[10] Indrayan A. Towards developing a health atlas of India: An exercise in health cartography. *Health and Population-Perspectives & Issues* 1988; **11:** 212–23.

[11] Weiss K B, Wagener D K. Geographic variations in US asthma mortality: Small-area analyses of excess mortality, 1981–85. *Am J Epidemiol* 1990; **132 (Suppl.):** S107–15.

[12] UNICEF. *An Analysis of the Situation of Children in India*. New Delhi; UNICEF Regional Office for South Central Asia, 1984.

[13] Raza M, Nangia S. *Atlas of the Child in India*. New Delhi: Concept Publishing Co., 1986.

[14] Bansal A K, Indrayan A. Computer based statistical study of cartography in mortality up to age of one year. *Indian Pediatr* 1993; **30:** 1251–58.

[15] Walter S D. The analysis of regional patterns in health data I. Distributional considerations. *Am J Epidemiol* 1992; **136:** 730–41.

[16] Clayton D G, Bernardinelli L, Montomoli C. Spatial correlation in ecological studies. *Int J Epidemiol* 1993; **22:** 1193–1202.

[17] Walter S D, Birnie S E. Mapping mortality and morbidity patterns: An international comparison. *Int J Epidemiol* 1991; **20:** 678–89.

[18] Anderberg M R. *Cluster Analysis for Applications*. New York: Academic Press, 1973.

[19] Jain A K, Dubes R C. *Algorithm for Clustering Data*. Englewood Cliffs: Prentice Hall, 1988.

[20] Smith S P, Dubes R. Stability of a hierarchical clustering. *Pattern Recog* 1980; **12:** 177–87.

[21] Bezdek J C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press, 1981.

[22] Thielemans A, Hoppe P K, De Quint P, Depoorter A M, Thiers G, Massart D L. Investigation of geographical distribution of female cancer patients in Belgium using pattern recognition techniques. *Int J Epidemiol* 1988; **17:** 724–31.

[23] Golden R R, Meehl P E. Detection of biological sex – an empirical test of cluster methods. *Mult Behav Res* 1980; **15:** 475–96.

[24] Jain N C, Indrayan A, Goel L R. Monte Carlo comparison of six hierarchical clustering methods on random data. *Pattern Recog* 1986; **19:** 95–99.

[25] Milligan G W, Schilling D A. Asymptotic and finite-sample characteristics of four external criterion measures. *Mult Behav Res* 1985; **20:** 97–109.

[26] Mezzich J E, Solomon H. *Taxonomy and Behavioral Science: Comparative Performance of Grouping Methods*. New York: Academic Press, 1980.

[27] Bayne C K, Beauchamp J J, Begovich C L, Kane V E. Monte Carlo comparisons of selected clustering procedures. *Pattern Recog* 1980; **12:** 51–62.

[28] Kuiper F K, Fisher L. A Monte Carlo comparison of six clustering procedures. *Biometrics* 1975; **31:** 777–83.

[29] SAS. *SAS/STAT User's Guide – Release 6.03 Edition*. Cary: SAS Institute Inc., 1988.

[30] SPSS. *SPSS/PC+ Professional Statistics Version 5.0*. Chicago: SPSS Inc., 1992.

[31] Cormack R M. A review of classification. *J R Stat Soc A* 1971; **134:** 321–67

[32] SRS. *Sample Registration System 1990*. New Delhi: Office of the Registrar-General, 1993.

[33] Verhasselt Y, Mansourian B. Methods for the classifications of countries according to health related indicators. *Bull World Health Organ* 1989; **67:** 81–84.

[34] Stanfel L E. Applications of cluster theory to cancer mortality data. *Comp Biomed Res* 1986; **19:** 117–41.

(Revised version received June 1995)