Pergamon

European Journal of Cancer Vol. 32A, No. 9, pp. 1551–1559, 1996

Printed in Great Britain
0959–8049/96 $15.00+0.00

PII: S0959-8049(96)00130-X

Original Paper

An Analysis of Alternative Classification Schemes for Medical Atlas Mapping

E.K. Cromley and R.G. Cromley

Department of Geography, University of Connecticut, 354 Mansfield Road, Storrs, Connecticut 06269-2148, U.S.A.

Most national disease atlases adopt a classification scheme based on either the percentile distribution of rates or on the national mean. Although these schemes have a direct interpretation, they are based on the univariate statistical distribution of rates and not on their spatial distribution, and distort the underlying spatial autocorrelation in the data. If the purpose of the maps is to represent spatial patterns, alternative classification schemes might be more appropriate. This research proposes an alternative classification method that maximises spatial similarity among contiguous units in the same class interval. The method has been illustrated using selected data from the German Cancer Atlas published in 1984. Copyright © 1996 Elsevier Science Ltd

Key words: medical atlas, cancer mapping, choropleth mapping, data classification
Eur J Cancer, Vol. 32A, No. 9, pp. 1551–1559, 1996

INTRODUCTION

MAPPING THE distribution of health problems has been undertaken to describe geographical patterns of disease for planning and resource allocation, and to identify clusters of disease as a preliminary step in explaining disease aetiology and transmission. The development of computer-assisted cartography and geographic information systems (GIS) has facilitated the production of these maps, making disease mapping available to any health specialist. However, because the ability to produce maps using computers is so common and poorly designed maps can inadvertently miscommunicate information [1], the need for improved mapping techniques is more important than ever. This paper considers an alternative to the classification schemes most commonly used in medical atlases. The classification scheme presented here explicitly uses the underlying spatial structure of disease rates to determine class interval limits.

For ease of presentation and interpretation, many medical atlases adopt a percentile classification scheme, often based on quintiles or sextiles [1–5]. In other atlases, only the extreme cases or those differing significantly from national norms are mapped [6, 7]. These classification schemes are not designed to represent statistically homogeneous groups or to highlight spatial patterns. Inferences from such maps, therefore, need to be conditioned by the limitations of these classification schemes. More recently, Becker [8] has proposed an absolute scaling technique that overcomes comparability problems between maps, but does not take into account the spatial structure of the disease distribution.

Medical statistics can be analysed using any number of techniques, but, to explore the spatial distribution of health problems, the statistics are generally presented in the form of a map. For statistics collected and reported by political/administrative units, these values are usually represented as a choropleth map. The display of medical statistics in the form of a chloropleth map first requires that the data be classified into a small number of class intervals, and then that a graphic value be assigned to represent the statistical values of all area units belonging to the same class interval. The graphic value is usually either a colour hue or a gray tone value.

Classification has been adopted in choropleth mapping, in part, to reduce the number of values being displayed because the human eye can only discriminate a limited number of tones [9], although perceptual studies conclude that the eye can discriminate more colour hues than gray tones [10]. Some researchers have proposed an alternative to classification in which each statistical value is assigned a unique graphic value and portrayed directly in a so-called 'classless' map [11–13]. This approach is based on the direct conversion of each statistical value into a unique graphic value.

The classification process transforms the original interval or ratio data into ordinal groups for display in the final map. A map reader can ascertain from the graphic values on the map

Correspondence to E.K. Cromley.
Received 3 Jan. 1996; accepted 27 Feb. 1996.

only that observations in one class are greater or less than observations in another class. The graphic values representing the statistical values are chosen either for maximum contrast between classes or are related to the mean or median numerical value for each class interval. A map legend is used to indicate the range of statistical values in the set of observations in each class.

Without careful study of the legend, comparisons between maps can be misleading. If graphic values were chosen for maximum contrast in each map, a dark tone on one map may represent a different range of statistical values than on another map. A quick visual comparison could lead the reader to misinterpretations of the relationships between the two disease distributions, a problem pointed out by Becker [8].

Even with careful study of the legend, the reader cannot determine how much individual statistical values within a class differ from one another. All metric properties of the statistical distribution are sacrificed for clarity of examining the spatial distribution. Therefore, the classification procedure should ensure that the information added by the graphic display more than compensates for the information lost by transforming individual statistical values into graphic value classes.

An alternative classification scheme, equal interval, reduces the variability of data ranges between groups. In this method, the data range of statistical values is divided by the number of classes. Each interval now has the same length, but the number of observations in each interval will vary. It is also possible that some class intervals will have no observations because no statistical properties of the distribution other than the range of the data were used. An equal interval scheme is useful for comparing maps if a common data range is used for all maps because the graphic value for each interval will be the same across all maps.

Serial class intervals [14, 15] represent a merging of the percentile and the equal interval methods. In serial classification, the full range of data values is divided into interval lengths based on a data transformation of the original statistical values. The type of transformation used is a function of the underlying mathematical relationship between a data value and its rank. Frequently, serial class intervals are based on a form of arithmetic, geometric or reciprocal hyperbolic progressions of the original statistical data [14]. Theoretically, in a serial classification, the same number of transformed data values will occur in each class interval.

A serial classification based on a linear data progression is the same as an equal interval classification. Becker [8] has proposed using a square root transformation of cancer mortality rates to create a 20 interval classification of cancer rates. A standard range of rates values from 0.0 to 95.0 is used to fix the intervals for the 20 class intervals across any cancer. By assigning a progression of 20 colour hues or gray tones to match the range of the 20 statistical intervals, maps having different data ranges can be compared because the graphic value always represents the same statistical interval.

These traditional methods have the advantage of being easy to compute and understand. However, they filter underlying spatial patterns of disease because the values of observations within an interval often vary widely resulting in a misrepresentation of disease clusters. In any serial classification based on a non-linear transformation, the width of each interval increases as the data values increase, increasing the likelihood of greater within-group variation.

To improve the value estimation in choropleth mapping, optimal classification strategies have been proposed that minimise the variation within classes [16–18]. Mak and Coulson [19] found that classed choropleth maps using an optimal classification system were significantly better for the task of value estimation than either n-tile or classless maps. Again, however, the statistical properties of the data distribution are not necessarily of primary concern in mapping because these properties can be analysed more appropriately by other techniques. If the primary concern is understanding the univariate distribution of rates, a frequency diagram is appropriate and most atlases include these graphics. If the primary concern is value estimation, a table of values by observational unit is appropriate, and some atlases include these tables. Maps are perhaps most useful for modelling the spatial properties of the data distribution. These properties of data distribution consider not only the data values but also the spatial contiguity of those values [20]. In this paper, an alternative classification scheme is presented.

MATERIALS AND METHODS

Classification based on the spatial arrangement of data values considers the amount of 'boundary error' that is present in a map after classification. The notion of boundary error is somewhat different from that of statistical error. Boundary error in choropleth mapping, a concept first introduced by Jenks and Caspall [21], occurs whenever the boundaries between the classed areas, referred to as external class boundaries, on the map do not align with the major breaks in a three-dimensional representation of the statistical surface. Classification should result in the boundaries lying within a group of contiguous area units, referred to as internal class boundaries, corresponding to minor breaks in the surface while the bounaries separating a grouping correspond to the major breaks in the surface. In this manner, contiguous area units belonging to the same class would form a statistically homogeneous cluster.

Boundary error is related to the type of spatial autocorrelation that is present in the data's spatial distribution. A number of significant breaks in a surface are present if the data are negatively autocorrelated. In this case, few if any clusters of similar values should occur as the result of the classification process. Fewer and less significant breaks occur if the data are positively autocorrelated. The classification process should identify a number of homogeneous clusters in this case. In practice, most phenomena exhibit moderate positive autocorrelation in space [22].

If the map is to portray regions that are homogeneous rather than heterogeneous, then the classification should try to minimise the size of the statistical breaks associated with internal class boundaries. This can be operationalised by calculating the level of similarity between area units belonging to the same class interval that share a common boundary. This similarity measure is added to a cost function for that class for all of its internal boundaries. Similarity measures associated with external values are ignored. A frequently used similarity measure is the square of the difference between the statistical values of the two area units associated with an internal boundary; this measure was first introduced by Geary [23] to calculate contiguity ratios. This measure is expressed as:

$$C_I = (X_r - X_l)^2 \quad (1)$$

where, C_I is the cost value and X_r and X_l are the statistical

Table 1. *Moran I coefficients for each cancer distribution*

Type of cancer	Moran I coefficient	Rank
Stomach		
Male	0.64	3
Female	0.76	1
Lung		
Male	0.73	2
Female	0.30	5
Colon		
Male	0.10	6
Female	0.32	4
Ovarian	0.07	7

values for the respective right and left area units for a given internal boundary. A dynamic programming algorithm for solving the problem of minimising this type of classification error over all classes has been presented by Cromley [18].

The alternative classification scheme employed here then classifies the units of observation to minimise the boundary error associated with the difference between the observed values squared for adjacent units of observation. In doing so, this classification scheme attempts to align large differences in the observed values with boundaries between classes and small differences with internal boundaries. To implement this type of classification scheme, not only are the statistical values for each area unit required, but also a topological structure identifying boundaries and their adjacent left and right area units (see Burrough [24] for a discussion of spatial data structures for storing digital maps).

RESULTS

To test this method as an alternative classification scheme, the technique was used to reclassify and map cancer mortality data from West Germany originally published and mapped in *Atlas of Cancer Mortality in the Federal Republic of Germany* [3]. The data in this atlas were collected at the level of the 'kreise' administrative unit and a mortality rate was estimated for each kreise for each cancer site by sex by the authors. In West Germany, there were 329 of these administrative units. For this study, the kreise of West Berlin was omitted because it was a detached unit at the time of the study that did not share a boundary with any other unit. In the atlas, all maps were published using a quintile classification of cancer rates.

The alternative classification proposed here was performed for lung cancer, stomach cancer and colon cancer rates for men and for women, and for ovarian cancer rates. These cancers were chosen from among more than 20 cancers in the original atlas because they have different underlying patterns of spatial autocorrelation. The Moran I spatial autocorrelation coefficient [25] was used to measure the amount and type of spatial autocorrelation for each cancer (Table 1). Male and female stomach cancers and male lung cancer exhibit the highest positive spatial autocorrelation ranging from 0.76 (for female stomach cancers) to 0.64 (for male stomach cancers). Female lung cancers and female colon cancers exhibited mod-

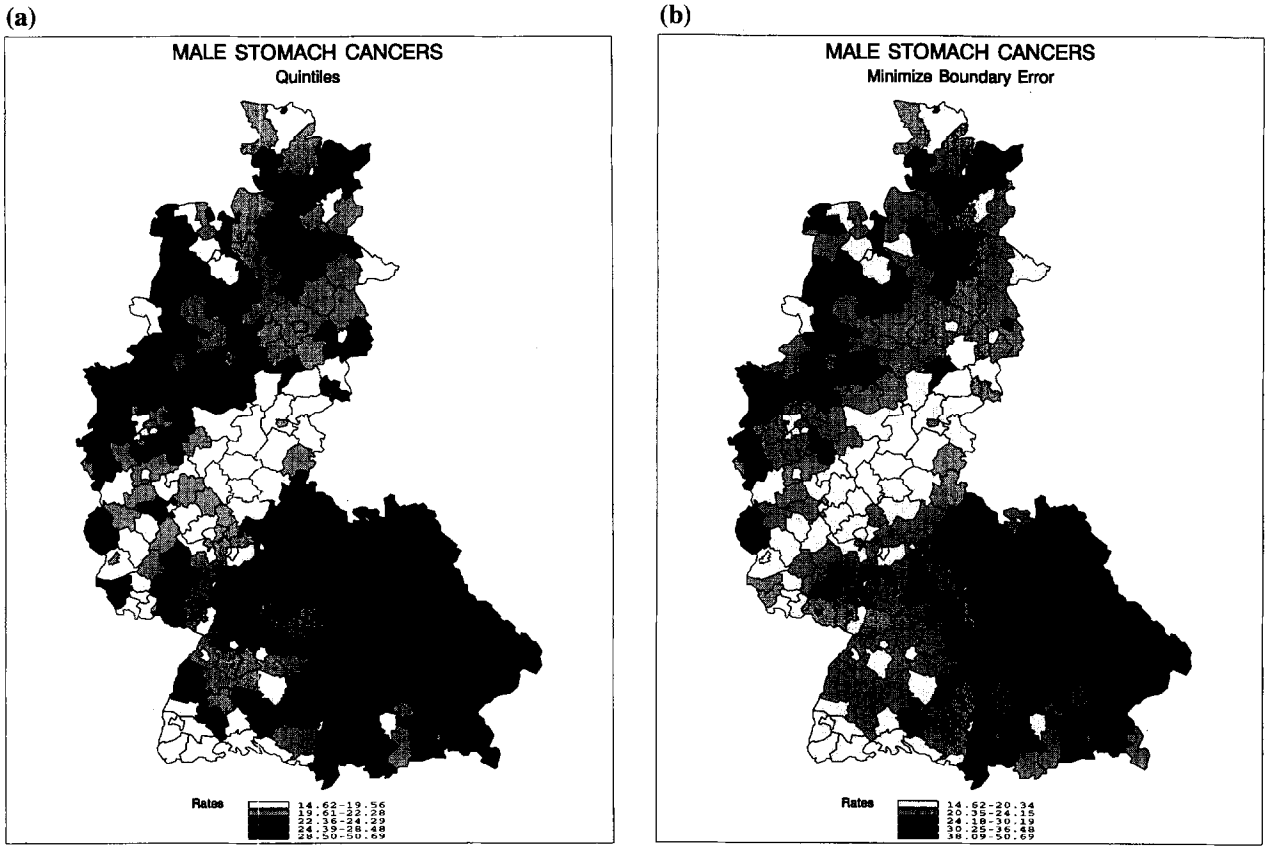

Figure 1. A comparison of male stomach cancers for West Germany. (a) Quintile classification; (b) minimise boundary error classification.

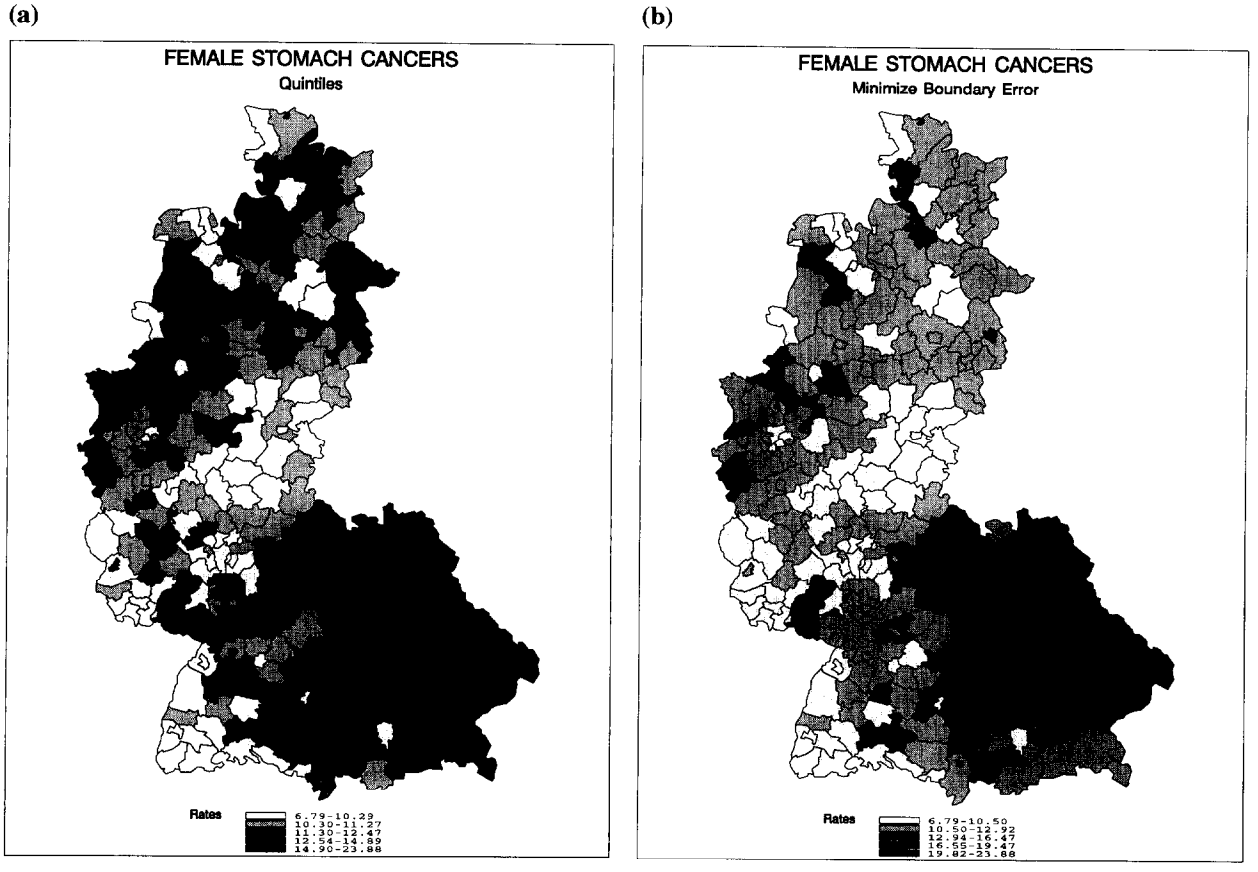

Figure 2. A comparison of female stomach cancers for West Germany. (a) Quintile classification; (b) minimise boundary error classification.

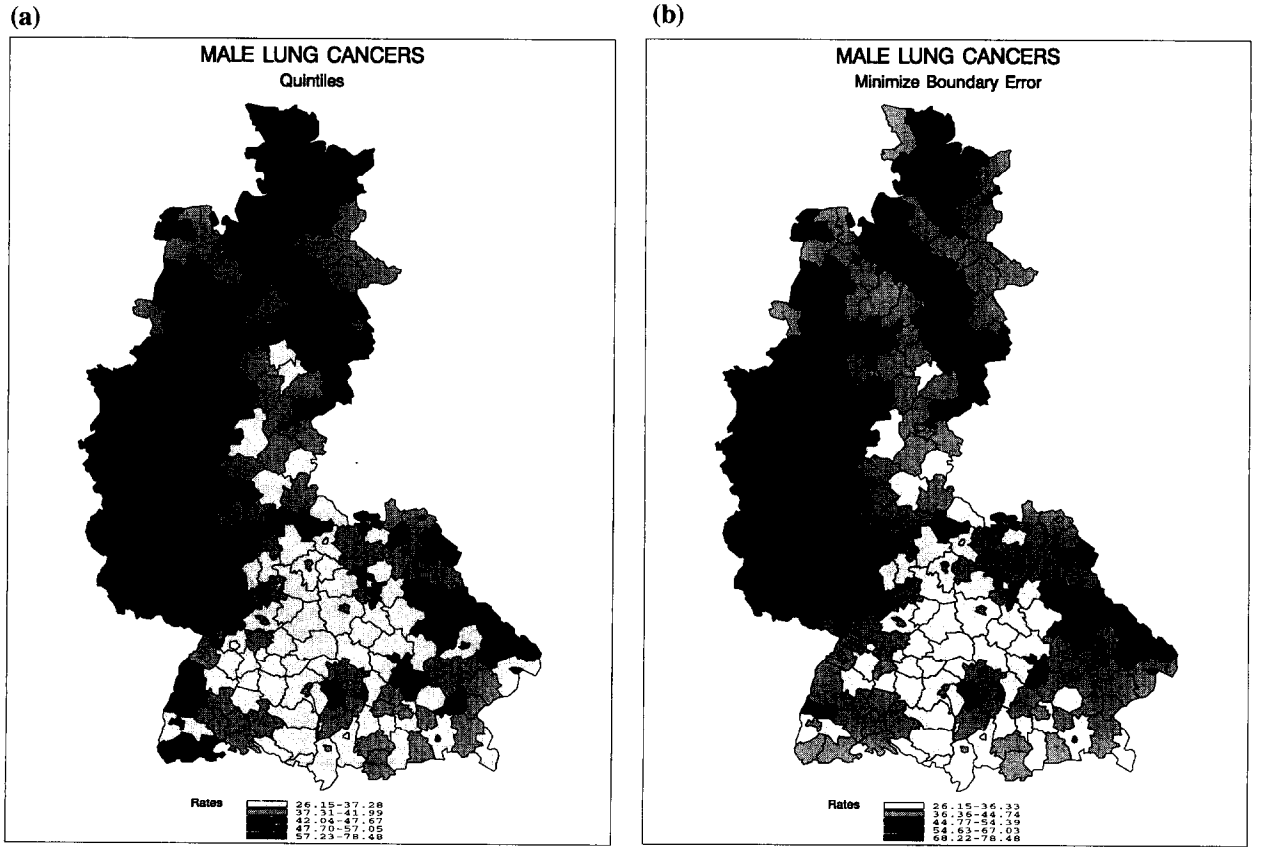

Figure 3. A comparison of male lung cancers for West Germany. (a) Quintile classification; (b) minimise boundary error classification.

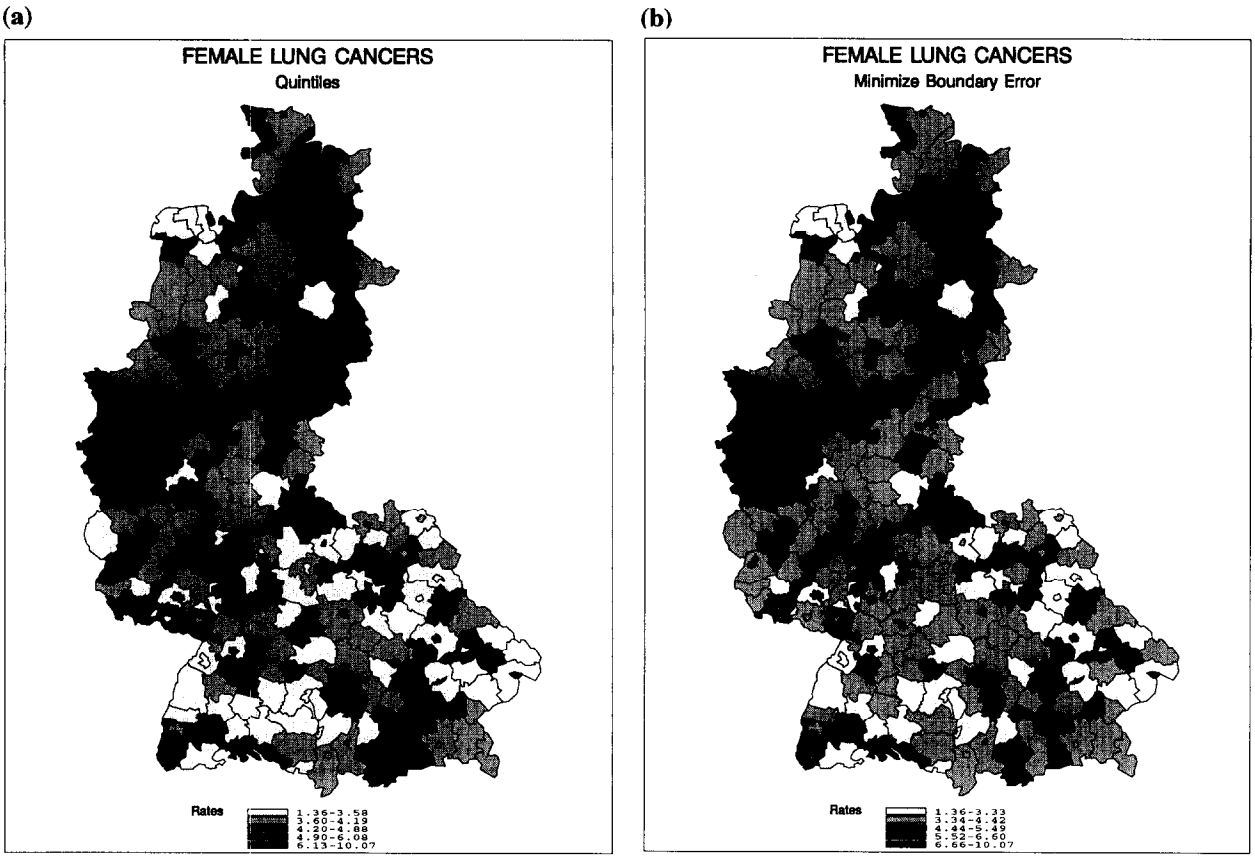

Figure 4. A comparison of female lung cancers for West Germany. (a) Quintile classification; (b) minimise boundary error classification.

erate positive spatial autocorrelation (around 0.30) and male colon and ovarian cancers were randomly distributed.

Using a quintile classification, maps of male stomach (Figure 1a), female stomach cancers (Figure 2a) and male lung cancer (Figure 3a) show a strong pattern of positive autocorrelation. For stomach cancer, a high concentration of cancer rates is observable in Bavaria in the southeastern portion of West Germany for both men and women. For lung cancer, among men, the highest category is concentrated in the heavily industrialised region of the Rhine and Ruhr River valleys in the west. Among female lung cancers, the kreises in the highest category are more dispersed and the area of high cancer rates in the west is substantially smaller (see Figure 4a). Female lung cancer does not appear to be as positively autocorrelated. Similarly, colon and ovarian cancers show randomness or negative autocorrelation (Figures 5a, 6a and 7a), although higher rates are observable in the western regions of West Germany for colon cancers (see Figures 5a and 6a).

These visual impressions of autocorrelation patterns associated with a quintile classification are also quantitatively measured by the number of internal and external class boundaries present in each map. As discussed above, a map should have a visual impression of positive autocorrelation if there are a large number of internal boundaries present in the display. According to this simple measure of spatial autocorrelation in map classification (Table 2), male lung cancers were the most positively autocorrelated followed by female and male stomach cancers, whereas the Moran I index ranked female stomach as being the most positively autocorrelated followed by male lung and male stomach. The remaining cancers have substantially fewer internal boundaries in their map pattern, with ovarian cancer having the least, although there was a reversal in the ranking for female lung and male colon.

The original data were then reclassified using the method described in the previous section that minimised the amount of boundary error. The maps resulting from this classification are juxtaposed with the quintile maps for visual comparison in Figures 1b–7b. For this classification, the rank order of the number of internal boundaries exactly matches the rank order of the Moran I index of spatial autocorrelation.

Because there is no longer a restriction that each class interval must contain a certain number of observations, the number of kreise associated with each class interval varied from class to class (see Table 3) better reflecting the underlying statistical and spatial properties of the data distribution. For male stomach cancers, the number of kreise in the highest category was reduced substantially from 66 to 8. In terms of the spatial distribution, the large homogeneous region of high cancer rates in southeastern West Germany is replaced by a somewhat more heterogeneous pattern while in the north, the more heterogeneous pattern in the quintile map is replaced by a more homogeneous pattern (see Figure 1). The same changes in statistical and spatial relationships were also present for female stomach cancers (see Figure 2). For male lung cancers, there are fewer very high cancer rates in the reclassed map (Figure 3b) than in the original quintile map (Figure 3a). The overall pattern has a smoother appearance with no units in southern West Germany in the upper two class intervals.

In contrast, for female lung cancers (Figures 4a and 4b), the quintile and reclassed maps are similar displaying a fair amount of spatial heterogeneity. This trend continued for both the colon and ovarian cancers (Figures 5, 6, and 7)—the

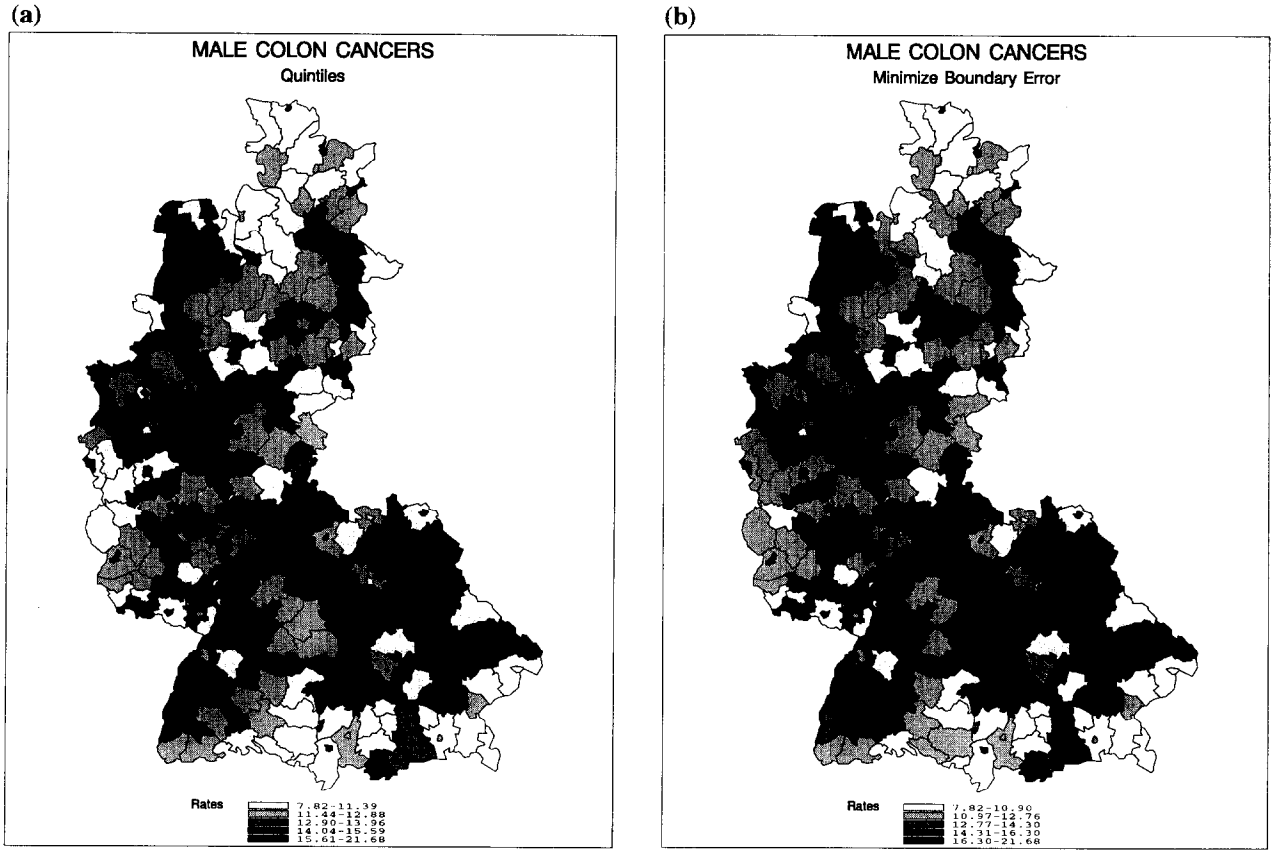

Figure 5. A comparison of male colon cancers for West Germany. (a) Quintile classification; (b) minimise boundary error classification.

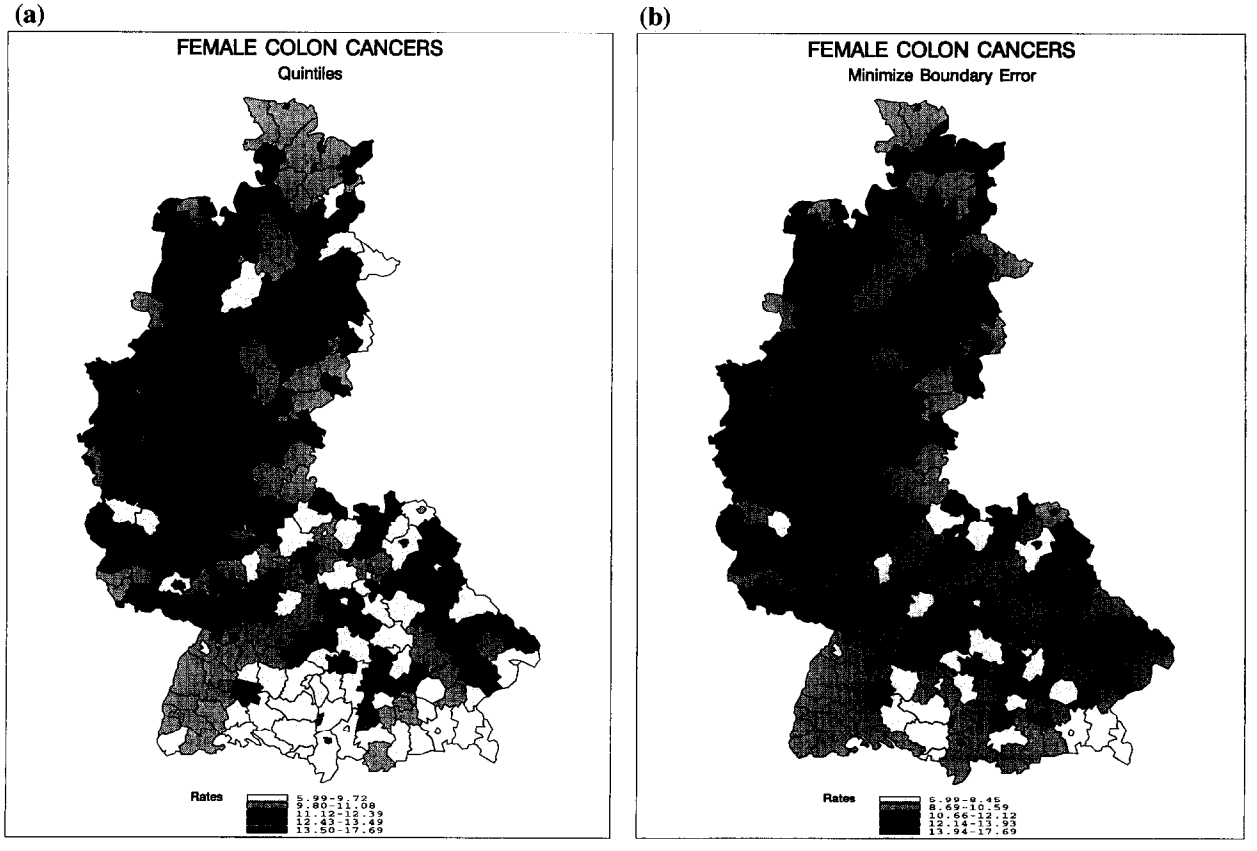

Figure 6. A comparison of female colon cancers for West Germany. (a) Quintile classification; (b) minimise boundary error classification.

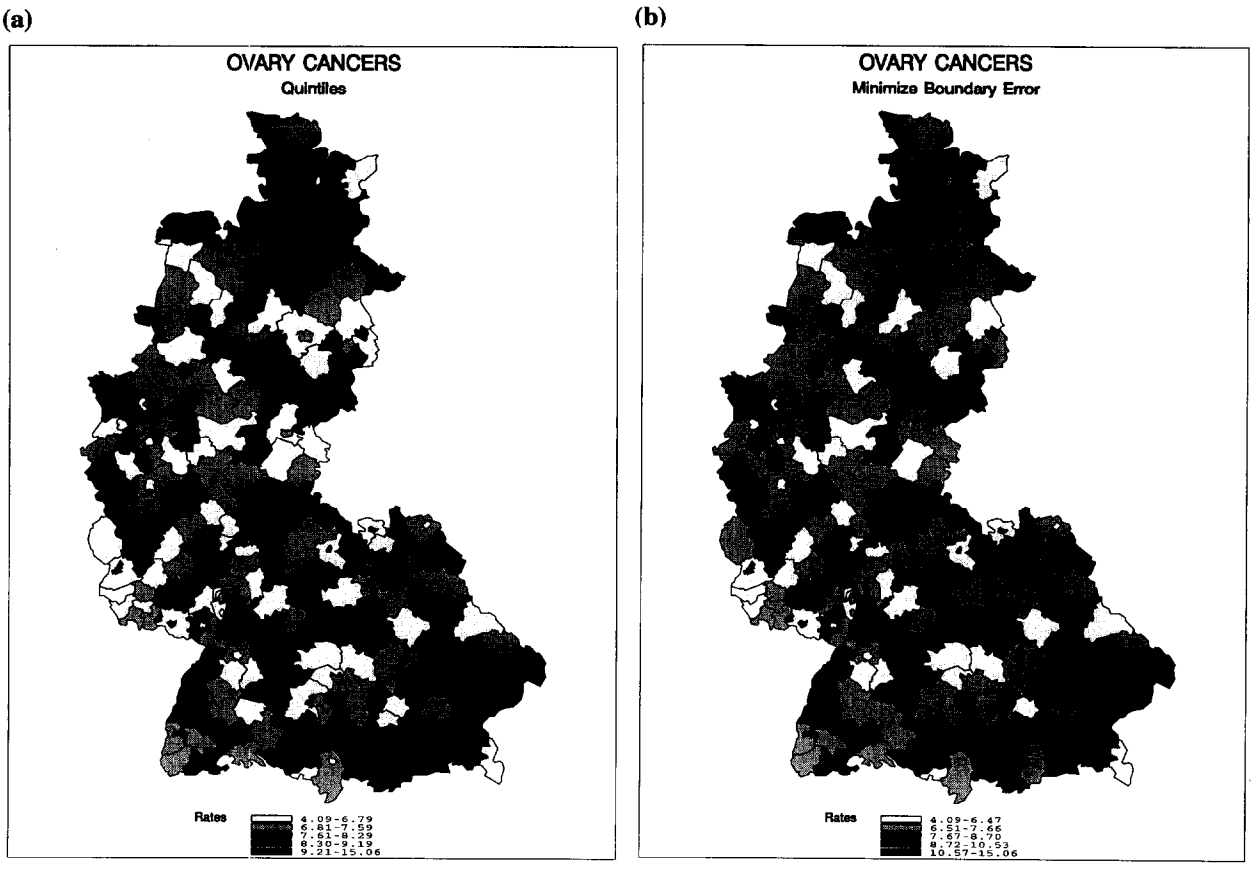

Figure 7. A comparison of ovarian cancers for West Germany. (a) Quintile classification; (b) minimise boundary error classification.

Table 2. Number of internal and external class boundaries associated with each classification scheme

Type of cancer	Number of internal class boundaries	Number of external class boundaries	Rank
Quintile classification			
Stomach, Male	318	527	3
Stomach, Female	328	517	2
Lung, Male	372	473	1
Lung, Female	222	623	6
Colon, Male	253	592	5
Colon, Female	268	577	4
Ovarian	182	663	7
Minimum boundary error classification			
Stomach, Male	347	498	3
Stomach, Female	385	460	1
Lung, Male	377	468	2
Lung, Female	234	611	5
Colon, Male	216	629	6
Colon, Female	253	592	4
Ovarian	197	648	7

overall visual impression of pattern is very similar and the number of observations within the five intervals is more uniformly distributed than for stomach or lung cancers (Table 3). As a further comparison of the two approaches, a statistical analysis was performed to determine how well each classification represented the underlying spatial structure of the thematic distribution. As spatial autocorrelation measures indicate, the squared observed value difference across an internal boundary should be small while the squared observed value difference across an external boundary should be larger.

Discriminant analyses were performed for each cancer to determine how well the observed difference, squared for each boundary, predicted whether a boundary would be internal or external under the two different classification schemes. In every case, the means of the observed differences squared for internal and external boundaries were a greater distance apart

Table 3. Number of observations associated with class interval for minimising boundary error

Type of cancer	Interval 1	2	3	4	5
Stomach					
Male	8	46	82	117	74
Female	15	30	68	138	77
Lung					
Male	17	60	94	99	56
Female	34	60	77	105	52
Colon					
Male	45	76	85	69	53
Female	44	97	81	79	27
Ovarian	28	66	96	94	44

the map reader with respect to the spatial properties of the disease under examination. Ease of diagnosis is of no benefit to medical researchers if the diagnosis is more likely to be in error.

An alternative classification is proposed here if the purpose of mapping is to examine pattern and identify clusters of areas having similar health statistics. A classification scheme that minimises boundary error in classified maps depicts such clusters where they exist more accurately than quintile maps. However, in the examples presented here, both classification schemes produced similar visual patterns when the data were negatively autocorrelated, although the internal and external boundaries had better statistical discriminatory power in the boundary error method. Hopefully, future use of the boundary error classification scheme will improve the utility of choroplethic mapping to medical researchers.

Table 4. Discriminant analysis associated with internal and external boundary categories

Type of cancer	Internal boundaries: Mean squared difference	Internal boundaries: Percentage correctly predicted	External boundaries: Mean squared difference	External boundaries: Percentage correctly predicted
Quintile classification				
Stomach, Male	14.3	87.7%	27.2	38.5%
Stomach, Female	3.1	83.5%	7.5	36.2%
Lung, Male	26.6	86.0%	86.7	45.7%
Lung, Female	0.5	93.2%	3.7	48.5%
Colon, Male	1.1	99.2%	11.8	47.8%
Colon, Female	0.9	93.0%	7.3	48.5%
Ovarian	0.9	94.0%	5.5	46.2%
Minimum boundary error classification				
Stomach, Male	2.7	98.6%	36.0	50.4%
Stomach, Female	1.0	98.2%	9.7	50.2%
Lung, Male	12.1	96.8%	99.0	55.1%
Lung, Female	0.2	98.7%	3.9	52.2%
Colon, Male	0.7	96.8%	11.8	48.6%
Colon, Female	0.4	99.2%	7.3	52.2%
Ovarian	0.3	100.0%	5.8	50.5%

for the boundary error classification than they were for the quintile (see Table 4). In all but one instance, internal boundaries for male colon cancer, the boundary error classification produced internal and external boundaries that better predicted what type of boundary their observed difference squared value said they should be.

One should also note that internal boundaries were more correctly predicted than external boundaries over all cancers. This is to be expected because these boundaries might be between kreise belonging to the first and second classes or the first and fifth classes creating greater variations in their differences than what would be expected in the difference variations for internal boundaries.

DISCUSSION

The usefulness of choropleth maps of disease for representing the spatial distribution of disease can be enhanced by incorporating classification techniques that better capture the spatial structure of the disease distribution. Although commonly used approaches, such as a quintile classification, are easily understood by map readers, they can also misinform

1. Monmonier MS. *How to Lie with Maps*. Chicago, University of Chicago Press, 1991.
2. Riggan WB, Creason JP, Nelson WC, *et al. U.S. Cancer Mortality Rates and Trends, 1950–1979, Vol. IV: Maps*. Washington DC, U.S. Government Printing Office, 1987.
3. Becker N, Frentzel-Beyme R, Wagner G. *Atlas of Cancer Mortality in the Federal Republic of Germany*. Berlin, Springer-Verlag, 1984.
4. Zatowski W, Becker N. *Atlas of Cancer Mortality in Poland, 1975–1979*. Berlin, Springer-Verlag, 1988.
5. Holland WW. *European Community Atlas of 'Avoidable Death'*. Second Edition, Oxford, Oxford University Press, 1991.
6. Mason T, McKay F, Hoover R, Blot W, Fraumeni J Jr. *Atlas of Cancer Mortality for U.S. Counties: 1950–1969*. Washington DC, DHEW Publication No. (NIH) 75–780, 1975.
7. Pickle LW, Mason T, Howard N, Hoover R, Fraumeni J Jr. *Atlas of the U.S. Cancer Mortality Among Whites: 1950–1980*. Washington DC, DHHS Publication No. (NIH) 87–2900, 1987.
8. Becker N. Cancer mapping: why not use absolute scales? *Eur J Cancer* 1994, **30A**, 699–706.
9. Jenks GF, Knos DS. The use of shading patterns in graded series. *Ann Assoc Am Geographers* 1961, **51**, 316–334.
10. Gilmartin P, Shelton E. Choropleth maps on high resolution CRTs/the effects of number of classes and hue on communication. *Cartographica* 1989, **26**, 40–52.
11. Tobler WR. Choropleth maps without class intervals? *Geographical Analysis* 1973, **5**, 262–265.

12. Peterson MP. An evaluation of unclassed crossed-line choropleth mapping. *Am Cartographer* 1979, **6**, 21–38.
13. Kennedy S. Unclassed choropleth maps revisited/some guidelines for the construction of unclassed and classed choropleth maps. *Cartographica* 1994, **31**, 16–25.
14. Jenks GF, Coulson MRC. Class intervals for statistical maps. *International Yearbook of Cartography* 1963, **3**, 119–134.
15. Evans IS. The selection of class intervals. *Trans Inst Brit Geographers* 1977, **2**, 98–124.
16. Monmonier MS. Analogs between class-interval selection and location-allocation models. *Canadian Cartographer* 1973, **10**, 123–131.
17. Jenks GF. *Optimal Data Classification for Choropleth Maps.* Occasional Paper No. 2, Department of Geography, University of Kansas, 1977.
18. Cromley RG. Optimal classification of spatially aggregated data. Paper presented at the Annual Meeting of the Association of American Geographers, San Diego, California, 1992.
19. Mak K, Coulson MRC. Map-user response to computer-generated choropleth maps: comparative experiments in classification and symbolization. *Cartography and Geographic Information Systems* 1991, **18**, 109–124.
20. Monmonier MS. Contiguity-biased class-interval selection: a method for simplifying patterns on statistical maps. *Geographical Rev* 1972, **62**, 203–228.
21. Jenks GF, Caspall FC. Error on choroplethic maps: definition, measurement, reduction. *Ann Assoc Am Geographers* 1971, **61**, 217–244.
22. Griffith D. On grouping for maximum homogeneity. *J Am Statist Assoc* 1994, **53**, 789–798.
23. Geary RC. The contiguity ratio and statistical mapping. *The Incorporated Statistician* 1954, **5**, 115–141.
24. Burrough PA. *Principles of Geographical Information Systems for Land Resources Assessment.* Clarendon Press, Oxford, 1986.
25. Moran PA. Notes on continuous stochastic phenomena. *Biometrika* 1950, **37**, 17–23.