Patterns in Geographical Data via Maximum Entropy Classification

Vladimir Estivill-Castro
School of Information Systems,
Queensland University of Technology,
Brisbane, 4000, Australia.

vlad@icis.qut.edu.au
phone: (61-7) 3864-1944
fax: (61-7) 3864-1969

February 25, 1997

Abstract

Classification is a fundamental operation for data analysis and pattern discovery in Geographical Information Systems. It consists of forming clusters from a large set of data values for presentation in choroplethic maps. Strategies based on a measure of diversity of the values within a class provide a quantitative measure of the quality of the classification but finding the classes requires $\Theta(n^3)$ time, where n is the number of data values.

We propose the use of classification via the Maximum Entropy Discretization (MED) of the probability distribution that is being represented by the choroplethic map. MED is rigorously founded on information theory. We show that our approach results in a method for classification that allows the analyst to treat the data as univariate or as spatially related. Moreover, the algorithms requires only $O(n \log n)$ time, and this requirement reduces to $O(n)$ time if data is already sorted.
TOPICS: Knowledge Discovery, Algorithms, Classification.

1 Introduction

A Geographic Information System (GIS) is designed for the collection, storage and analysis of objects and phenomena where geographic location is an important characteristic or critical to the data analysis. The power of a GIS lies in its ability to analyze spatial and attribute data together. One of the important functions of a GIS is facilitating the recognition of patterns for decision making. It is this capability that most distinguishes a GIS from automated mapping and computer-aided draft systems.

The available functionalities in GIS technology for analysis procedures are very large. Among these, *classification* [1] is fundamental because it defines patterns. Classification is the procedure of identifying a set of features as belonging to a group. For example, consider the set of land parcels in an urban area and the age of the building in the parcel. The parcels could be assigned a class such as "pre-1900", "1900-1930', "1930-1945" and "post-1945"; then, presented colored according to the class they belong.

Classification is basically a cluster analysis problem [6]. We are given a set of values $X = \{x_1, x_2, \ldots, x_n\}$ and we want to find a set of intervals (also called clusters, classes or types) such that the values within one interval will be considered similar, or belonging to the same class or type. The interval provides a class description, and a prototypical point may be sometimes chosen as a representative of the class (for example, the mid-point of the interval).

At first, classification may seem to loose information; however, even in the extreme case that all data points are grouped in one class, we obtain the range of values (that is, additional order statistics). On the other hand, *generalization* [1], also called *map dissolve*, is the process of making classification less detailed by combining classes. For example, combining regions for Wetland-Treed Bog and Wetland-Open Bog into a more general class Wetland. Generalization is often used to reduce the level of classification to make the underlying pattern more apparent. Classification may be regarded as a generalization of the partition where each point is in a class by itself.

Although classification may or may not be guided by domain knowledge, classification is used to present summarized information to the human ability to detect patterns from the structure of the distribution of values in the set X. Also, that the classification reflects information of the distribution of values is important so further inferences have reasonable grounds. For example, if a map of our urban area was produced with the given labels, and it is believed that older buildings have a higher level of risk in earthquakes, the total area under risk could be approximated. In fact, every map is the result of classification to the pixel's level, since everything smaller than the size of the pixel belongs to the same class as the label of the pixel. Moreover, classification in the GIS is analogous to unsupervized classification in machine learning [7], and thus, it is a process of knowledge discovery.

What constitutes a good classification depends on the application at hand. Classification schemes for a GIS have been grouped in two broad categories [5]. An *exogenous* classification is determined by a priory knowledge without regard to the data distribution itself. An illustration of this could be the selection of the limits of the interval "1930-1945" on a priori knowledge of historical events. Alternatively, *ideographic* classification is a function of the data distribution to be displayed and its purpose is to reflect the underlying characteristics of the statistical distribution.

Since some form of classification function is provided in every GIS, and visualizing or printing a map is classification to the discrete values possible at pixel level, the premise of this article is that automated classification is obviously necessary and worthwhile.

Recently, several classification strategies were compared on empirical and real-world data [4]. There was not a clear winner and the most informative result was that no strategy could be considered inferior. Because each of these strategies minimizes a measure of the variability within a class, (with the side effect of maximizing the similarity between class elements) they are called optimization classification strategies [4]. Although these strategies reflect the nature of the statistical distribution, their computational complexity requirements grow faster than a small polynomial on the number of data values (the algorithms are $\Theta(n^3)$ time). Since GIS applications usually involve large data, algorithms for classification most deal with the issue of computational complexity.

We propose the use of the Maximum Entropy Discretization (MED) [9] for classification in GIS. MED is based on rigorous minimization of information loss by maximizing the entropy of the event covering produced by the classes. Moreover, this process allows the diversity (or difference) between classes to be maximized [2]. MED has been originally used in pattern-matching applications [9] and more recently, in knowledge discovery in data bases [2].

In this paper we demonstrate that MED is suitable for classification in GIS because

- it requires linear time on the number n of data points when these are already sorted; and $O(n \log n)$ time otherwise;
- the classification reflects the underlying characteristic of the statistical distribution as spatially observed;
- it is rigorously based on information theoretic measures by Shanon [11, 12];
- empirical illustration on simulated and real-data show its value for classification.

The paper is organized as follows. Section 2 presents Maximum Entropy Discretization. Section 3 presents our proposal to use MED with geographical data. Section 4 compares MED with optimization classification strategies. We conclude with some final remarks in Section 5.

2 Maximum Entropy Discretization

A fundamental purpose of mapping data spatially related is to clearly illustrate the patterns of areal variation and present a useful representation of the distribution. The many values in original data most be grouped into classes so that these may be symbolized, usually by areal shading patterns. Map-readers gain their impression of the distribution from whatever pattern is formed by the class intervals selected.

Classification into a relatively few categories (seven or eight is regarded as the maximum for human understanding [8]) can have many solutions [4]. We need to be able to compare and rank the selection of classes in a quantitative manner, using some measure of quality. The approach we propose here is to use the information content of the choroplethic map using ideas from information theory. On one hand, *choroplethic maps* [4] or *isometric statistical block diagrams* [8] are graphical simplifications for the histogram of the distribution that in turn is a graphical representation of a set of probabilities for an event-cover of the sample space of a probability distribution. On the other hand, *information theory* is a layer over probability theory that allows to compute information content on a set of events given their probabilities. Information theory has been the mathematical tool for many solutions to communications problems where typically we are interested in the efficient transmission and reception of information from point A to point B. We may consider drawing a choroplethic map as the latest phase of the transmission of an informative pattern to the map-reader (point B). Information transmission starts at the data collection phase (surveying and/or remote sensing) where initial signals about the environment (point A) are obtained. Although there are many differences in the communication of patterns in spatial data to GIS users and communications theory, this analogy is presented here as an indication of the role of information theory in pattern discovery.

We present classification by Maximum Entropy Discretization (MED) using an example and with the help of classification by Equal Interval Ranges (EIR).

Let X be a random variable and consider the following set of 30 observations of X: 0.1, 0.9, 1.5, 2.0, 2.8, 3.2, 3.3, 3.5, 3.7, 3.8, 4.0, 4.5, 4.9, 5.5, 6.0, 7.3, 8.5, 8.8, 9.1, 9.2, 9.5, 9.5, 9.7, 9.7, 10.0, 10.3, 10.5, 11.1, 11.8, 12.9. If we divide the range $[0.1, 12.9]$ of observed values into 5 equally sized intervals (EIR [4]) and use them as a cover for a histogram, a first visual picture of the probability density for X appears. Let I_i denote the i-th interval, for $i \in \{1, \ldots, 5\}$. The probability that X is in I_i is estimated (by the maximum likehood estimate) as the ratio of the number of observation in I_i over the total number of observations. The histogram represents the probability distribution function $\phi(X)$ for X as a constant for each interval I_i. Thus, for $i \in \{1, \ldots, 5\}$,

$$\begin{aligned} Prob(X \in I_i) &= \int_{I_i} \phi(X) dx \\ &\approx |I_i| height(I_i), \end{aligned}$$

where $|I_i|$ is the length of interval I_i and $height(I_i)$ is the height of the histogram in I_i. From this we obtain that

$$height(I_i) = \frac{\sharp \text{ of observations in } I_i}{(\sharp \text{ of observations })|I_i|}.$$

For our example, in the case of ranges of equal length, $|I_i| = 2.56$. Figure 1 (a) presents the corresponding illustration that should provide more information on the patterns of variation of X than the list of 30 data values.

Entropy is an information theoretic measure of randomness in a random variable. In general terms, the entropy $H(X)$ of a random variable X is defined as the sum

$$H(X) = -\sum_i q_i \log_2 q_i,$$

where q_i is the probability that X takes on the i-th value. Conventionally, logarithms are to the base 2, and entropy is then said to be measured in units called "bits" (binary information units). The special case of equal probabilities maximizes the entropy. If there are k possible

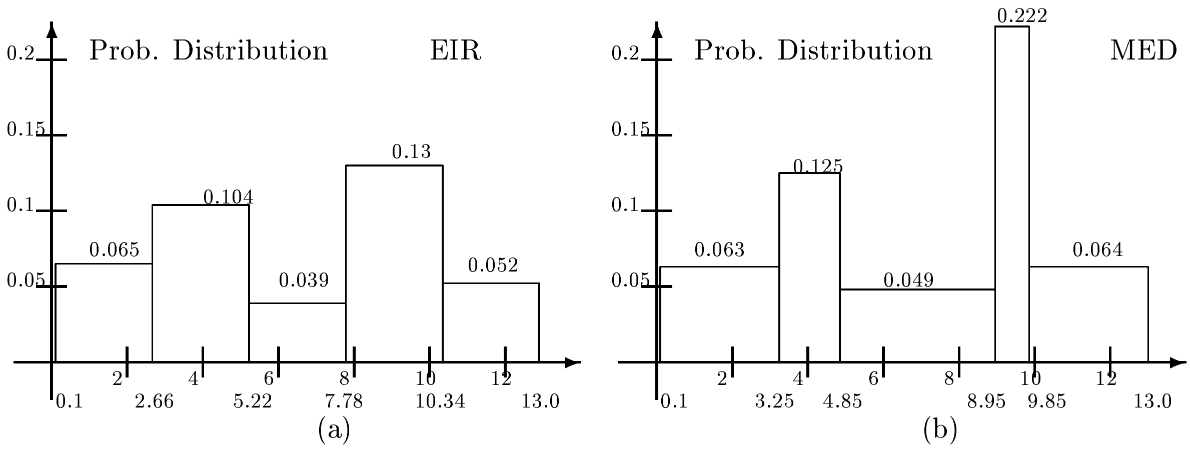

Figure 1: Histrograms generated as a result of calssification. Equalt Interval Ranges (EIR) results in part (a) while Maximum Entropy Discretization (MED) results in part (b).

events, the maximal entropy is $\log k$. In the context of classification, a point to note is that no variation is zero entropy, and thus there is no information for discrimination of classes.

Maximum Entropy Discretization (MED [9]) seeks equal probability. Therefore, selecting the number of classes as 5 implies finding 5 disjoint intervals ranges I_i for which

$$Prob(X \in I_i) = Prob(X \in I_j)\ \forall i, j \in \{1, \ldots, 5\}.$$

Using maximum likehood estimation, this means

$$\frac{\sharp \text{ of observations in } I_i}{(\sharp \text{ of observations })} = \frac{\sharp \text{ of observations in } I_j}{(\sharp \text{ of observations })}\ \forall i, j \in \{1, \ldots, 5\};$$

or

$$\sharp \text{ of observations in } I_i = \sharp \text{ of observations in } I_j,\ \forall i, j \in \{1, \ldots, 5\}.$$

That is, we need to find endpoints so that all intervals have the same number of data observations. In our example, any set of endpoints that places 30/5=6 observations per interval will maximize the entropy. Thus, in MED we have that, for all i,

$$\begin{aligned} heigh(I_i) &= \frac{Prob(X \in I_i)}{|I_i|} \\ &= \frac{\sharp \text{ of observations in } I_i}{(\sharp \text{ of observations})|I_i|} \\ &= \frac{(\sharp \text{ of observations})/(\sharp \text{ of intervals})}{(\sharp \text{ of observations})|I_i|} \\ &= \frac{1}{(\sharp \text{ of intervals})|I_i|}, \end{aligned}$$

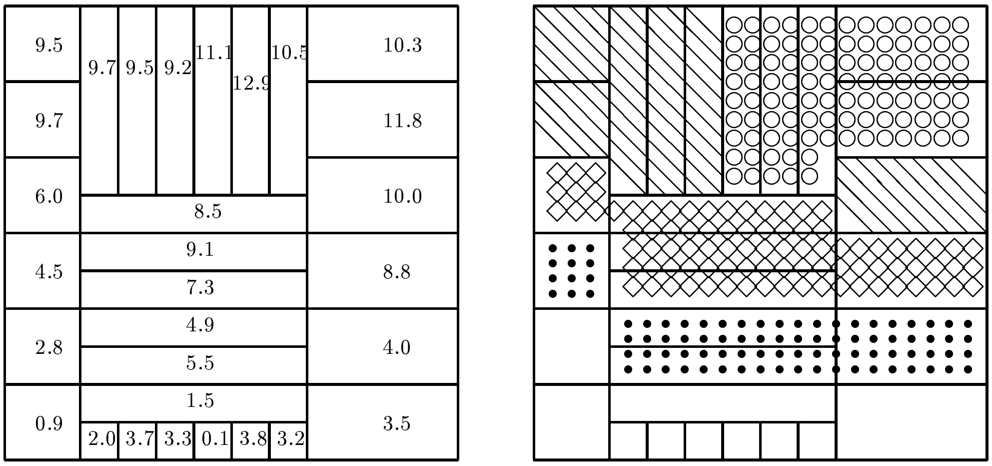

Figure 2: The X of 30 values spatially associated and the classification into 5 classes via MED with regard to area of associated regions.

where, again, $|I_i|$ is the length of interval I_i.

For example, Figure 1 (b) presents the corresponding histogram for Maximum Entropy Discretization. A remark should be made regarding EIR and MED. As illustrated by the comparison in Figure 1, MED shows clearly the peaks of the distribution and thus, the reader of the second histogram has a better picture of the variation of X across the domain of possible values.

3 Using MED for geographical data

Our proposal of MED is not only rigorously based on information theory (as illustrated in the previous section) but its application to spatial data naturally brings two possibilities for its use.

The first approach is to regard the numeric attribute-valued data as univariate data (with little regard to its association with a bi-dimensional geographical region that has some extension) [4, 10]. This was illustrated in the previous section (refer to Figure 1 (b)) where a set X of 30 values was clustered into 5 classes.

The second approach is to regard the attribute-valued data as the observation of bivariate data that is constant within the geographical region [3, 8]. For example, imagine that the example set X of observations corresponds to a region as illustrated in Figure 2. That is, within a region, the observed value of X is constant. Note the fact that the value 10.3 is observed over the North-most East-most region which has twice the area of the South-most West-most region (that has a value of 0.9) makes the difference between regions more significant. This could be reflected by MED classification. Thus, in this second approach, the histogram that approximates the distribution has a bi-dimensional domain (with an assumption of the statistical independence of the marginal distributions).

The algorithms for both approaches are formally presented now. The algorithm *MED-U* for classification of univariate data is presented in Figure 3. *Med-U* is conceptually very

Algorith MED-U

Inputs: 1) k (the number of classes),
2) X (the set of n data values).

Output: 1) A set of k disjoint intervals I_i covering the range of values of X such that:
$Prob(X \in I_i) = Prob(X \in I_j)\ \forall i, j \in \{1, \ldots, k\}$.

Step 1: Sort X to obtain the sorted sequence
$\langle x_1, x_2, \ldots, x_n \rangle$.

Step 2: Compute interval cardinality IC as:
$IC \leftarrow \lfloor n/k \rfloor$.

Step 3: Compute endpoints of classes:
$I_1 \leftarrow [x_1, (x_{IC-1} + x_{IC})/2\ ),$
$I_k \leftarrow [(x_{(k-1)(IC)-1} + x_{(k-1)(IC)})/2, , x_n + \epsilon),$
for $i \leftarrow 2$ **to** $k-1$
$I_i \leftarrow [(x_{(i-1)(IC)-1} + x_{(i-1)(IC)})/2, (x_{i(IC)-1} + x_{i(IC)})/2\ ).$

Figure 3: Algorithmic description of MED for univariate data.

simple. We need to find k intervals covering the range of n data values so that each interval covers the same number of values. Thus, once the values are sorted and we now how many values should be in each interval, we scan the sorted sequence of values grouping them in intervals of equal size. In Figure 3, *MED-U* computes the rightmost endpoint of interval I_i as the mid-point of the largest item in I_i and the smallest of the items that are larger than items in I_i. Similarly for the leftmost endpoint. Some care is required for the first and last intervals.

The algorithm *MED-R* for classification of data associated to regions is presented in Figure 4. To demonstrate the *MED-R* classification algorithm consider Figure 2 (a). The set X had its values already in sorted order and the total area of the map is 144 (square units). Assuming we are grouping into 5 classes, each class must cover about 144/5 or 28.8 (square units). Thus, we add the areas of regions according to the order in X until we cover 28.8 (square units). The area for the regions with X values 0.1, 0.9, 0.5, 2.0, 2.8, 3.2, 3.3, 3.5, 3.7 and 3.8 adds to 28 (square units) while if we also include the region with value 4.0 we get 30 (square units). Thus, the first interval (class) is $[0.1, (3.8 + 4.0)/2) = [0.1, 3.9)$. Now, we continue the scan of the area partial summation for the next break point at twice 28.8; that is 57.6 (square units). This occurs when x_1 to x_{14} are considered, since their area regions add to 52 (square units) but the area including also x_{15} results in 58 (square units). Thus,

$$I_2 = [3.9, (x_{14} + x_{15})/2) = [3.9, (5.5 + 6.0)/2) = [3.9, 5.55).$$

The final classification is shown in Figure 2 (b).

ALGORITH MED-R

Inputs: 1) k (the number of classes),
2) X (the set of n data values),
3) A (an area size for each data value),
4) R (a region for each data value).

Output: 1) A set of k disjoint intervals I_i covering the range of values of X such that if a point is uniformly selected on the union of the regions, then
$Prob(X \in I_i) = Prob(X \in I_j)\ \forall i, j \in \{1, \ldots, k\}$.

Step 1: Sort X to obtain the sorted sequence
$\langle x_1, x_2, \ldots, x_n \rangle$.

Step 2: Compute TA, the total area of the regions as:
$TA \leftarrow \sum_{i=1}^{n} A_i$.

Step 3: Compute AS, the area size for equi-probable sets as:
$AS \leftarrow \lfloor TA/k \rfloor$.

Step 4: Compute endpoints of area sizes:
for $i \leftarrow 1$ **to** k
t_i is such that
$\sum_{i=1}^{t_i} A_i \leq i(AS)$ and $\sum_{i=1}^{t_i+1} A_i > i(AS)$.

Step 5: Compute endpoints of intervals as
$I_1 \leftarrow [x_1, (x_{t_1} + x_{t_1+1})/2\ )$
$I_k \leftarrow [(x_{t_{k-1}} + x_{t_{k-1}+1})/2, x_n + \epsilon)$
for $i \leftarrow 2$ **to** $k-1$ **do**
$I_i \leftarrow [(x_{t_{i-1}} + x_{t_{i-1}+1})/2, (x_{t_i} + x_{t_i+1})/2\ )$.

Figure 4: Algorithmic description of MED for data associated to regions with area values.

The algorithms are very efficient. The most time consuming operation is the sorting of the values in X, which requires $O(n \log n)$ by comparison-based sorting methods. If the data are already sorted the algorithms require $O(n)$ time, essentially, the time to read the input.

4 A comparison with optimization strategies

Several qualitative measures of the quality of classification have been proposed [4]. These evaluate the quality of a classification by the variability within a class by several metrics. They find the grouping that minimizes this measure by solving and Integer Programming Optimization Problem [4]; thus, we refer to them as optimization strategies. In this section we compare MED with these strategies [4].

First we show that, in terms of computational resources, MED is a clear winner since it requires only $O(n \log n)$ time, where n is the number of data values. On the other hand, the metric-optimization strategies will require at least $\Omega(n^3)$ time.

Later, we will present a comparisons using real data of the classification performed by MED and the classification performed by the optimization strategies.

4.1 On the complexity issue

Consider the following metrics to evaluate the variability between a group of sorted values $G_i = \{x_{i_1}, x_{i_2}, \ldots x_{i_{g_i}}\}$ (where at least two are different) and a representative x for the group.

1. Variability L_2-based, defined by

$$Var_{L_2}(x, G_i) = [\sum_{j=1}^{g_i} (x - x_{i_j})^2]^{1/2}.$$

2. Variability L_1-based, defined by

$$Var_{L_1}(x, G_i) = \sum_{j=1}^{g_i} |x - x_{i_j}|.$$

3. Variability L_∞-based, defined by

$$Var_{L_\infty}(x, G_i) = \max_{j=1,\ldots,g_i} |x - x_{i_j}|.$$

In fact, from the perspective of the L_2 metric (the Euclidian world), the best representative of the group is the arithmetic mean; that is,

$$\hat{x} = \frac{1}{g_i} \sum_{j=1}^{g_i} x_{i_j},$$

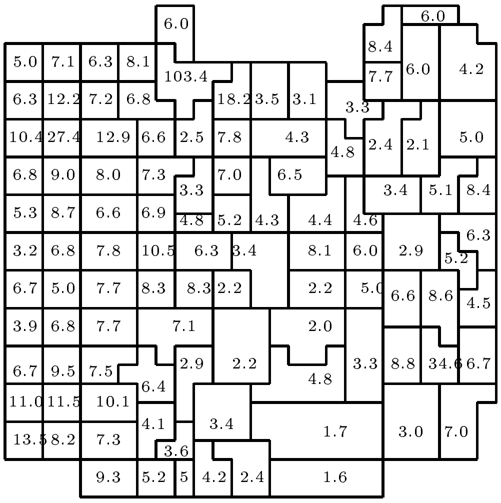

Figure 5: The original data set of rural population densities of 7 countries in central Kansas.

since $Var_{L_2}(\hat{(x)}, G_i)$ is minimum. From the perspective of the L_∞ norm, the midpoint of the range, (that is, $(x_{i_1} + x_{i_{g_i}})/2$) is the best representative. Finally, from the perspective of L_1 (the Manhattan norm), the median (that is, the mid-element $x_{i_{\lfloor g_i/2 \rfloor}}$)) is the best representative.

Thus, optimization classification strategies have two qualitative approaches:

- They find k groups G_i in such a way that the *total variation* in each group (and its best representative for the metric) is minimum. Namely, minimize

$$\sum_{i=1}^{k} Var(G_i).$$

- They find k groups G_i in such a way that the *largest of the variations* in each group (and its best representative for the metric) is minimum. Namely, minimize

$$\max_{i=1,\ldots,k} Var(G_i).$$

Since there are three metrics and two approaches, we have six optimization classification strategies.

To compute the optimum classification for either of this six criteria, the corresponding optimization problem is transformed into an acyclic network [4], and then into a integer programming problem.

Although integer programming is NP-hard in general, the fact that the network is acyclic means that this special case can be solved in polynomial time [4]. However, the best algorithm developed so far has an exponent large enough that the optimization problems is approximated by a Lagrangian relaxation [4].

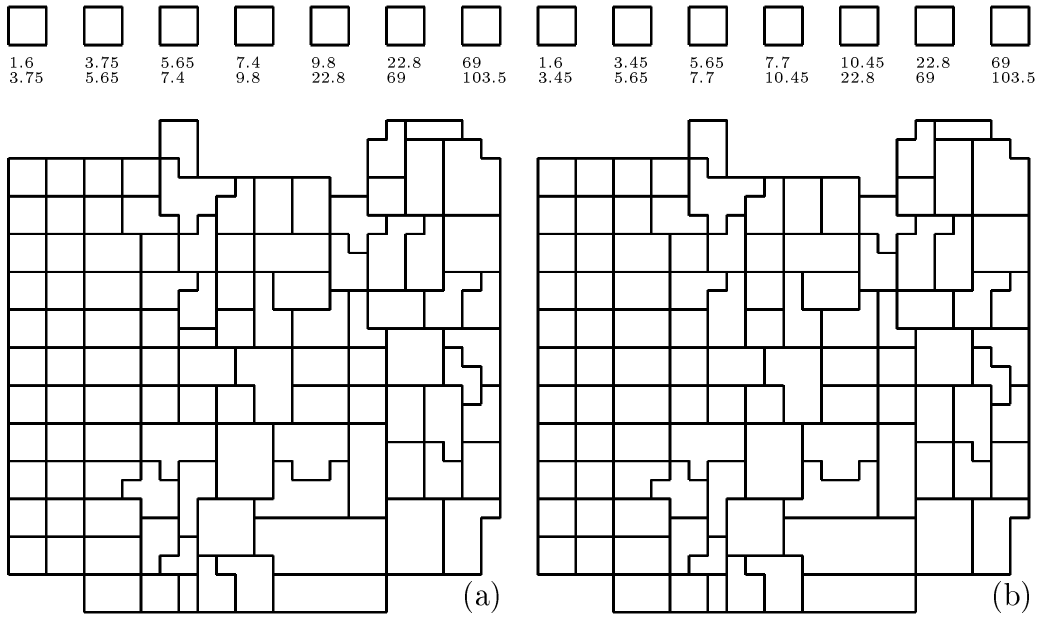

Figure 6: The L_1 classification maps. Map (a) uses minimization of total variability of groups while map (b) uses minimization of largest variability.

We show that the writing of the corresponding shortest path problem requires $\Theta(n^3)$ time. Therefore, no matter what optimization classification strategy is used and how its corresponding optimization problem is solved or approximated, these methods require $\Theta(n^3)$ time.

The problem of each of the optimization strategies is represented as a form of shortest path finding over an acyclic network as follows [4]. Let N be the set of $n+1$ nodes in the network. The nodes are labeled $\{1, 2, \ldots, n, n+1\}$. The arcs of the network are the ordered pairs (i, j) with $i < j$. Thus, there are $\Theta(n^2)$ directed arcs, and clearly, the network is acyclic since any path must visit nodes in ascending order. Paths from node 1 to node $n+1$ correspond to a classification. If arc (i, j) is included in a path, the classification has a group with items $\{x_i, x_{i+1}, \ldots, x_{j-1}\}$. Paths of length k correspond to classification into k classes.

The cost c_{ij} of the arc (i, j) is set according to the classification strategy that must be solved optimized. Thus, c_{ij} is the variability in group $\{x_i, x_{i+1}, \ldots, x_{j-1}\}$ with respect to its best representative. Thus, computing the $\Theta(n^2)$ costs of the network (or the $\Theta(n^2)$ cost values needed for the integer programming problem) requires $\Theta(n^3)$ time.

Similarly, there is a point to be amde about storage requirements. Our proposal of using MED requires only linear space, while the approach on optimization strategies requires $\Theta(n^2)$ space. This quadratic space is required even during the solution of the problem since the algorithms for optimization strategies are dynamic programming algorithms.

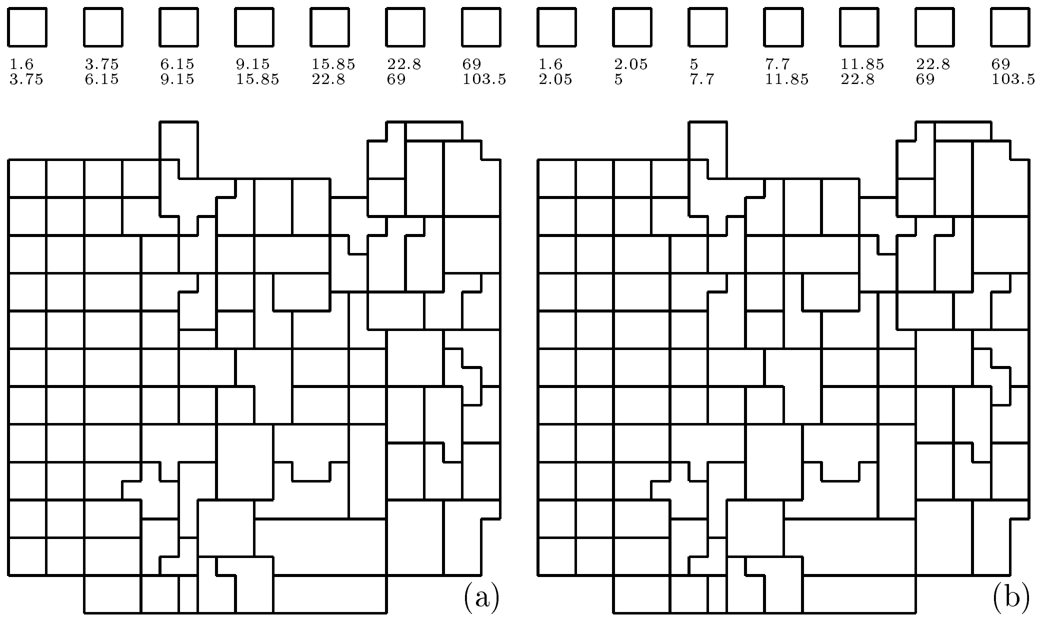

Figure 7: The L_2 classification maps. Map (a) uses minimization of total variability of groups while map (b) uses minimization of largest variability.

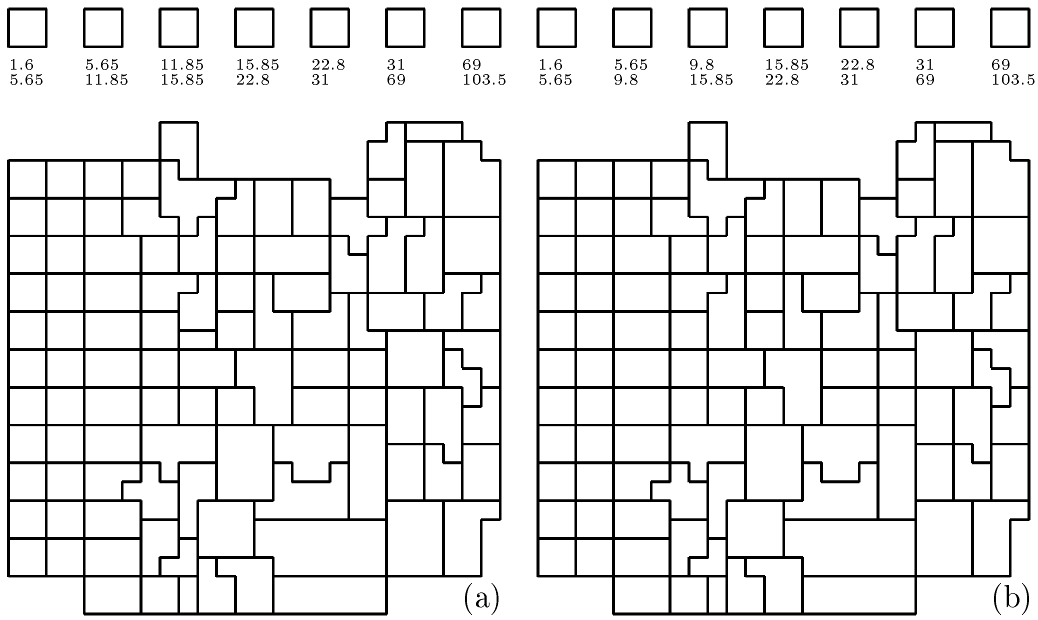

Figure 8: The L_∞ classification maps. Map (a) uses minimization of total variability of groups while map (b) uses minimization of largest variability.

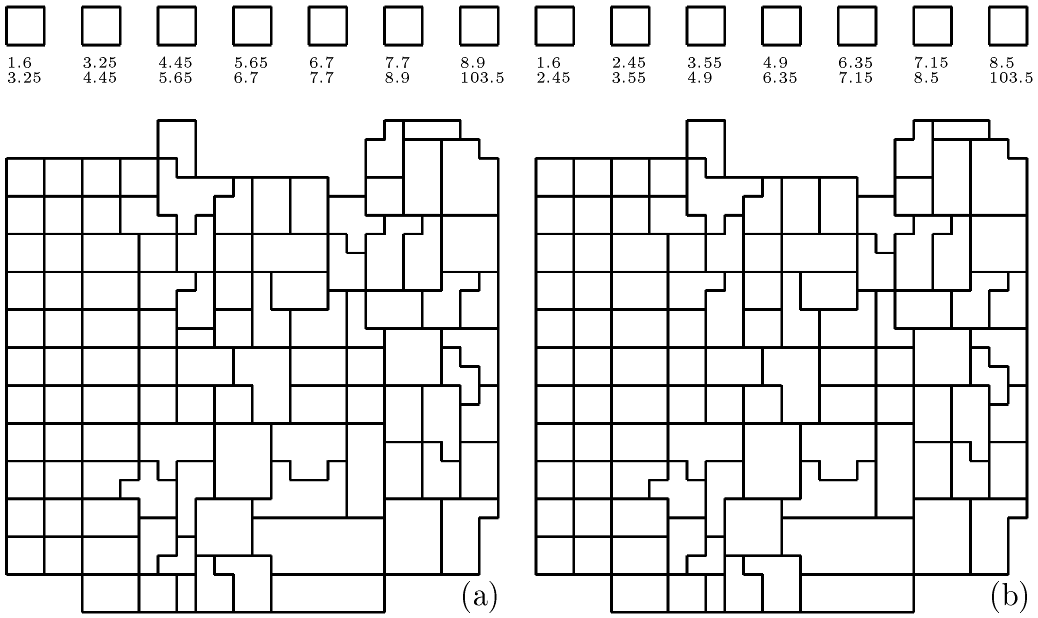

Figure 9: The MED classification maps. Map (a) uses the univariate version of MED while map (b) used the bi-dimensional version.

4.2 A comparison using real data

In this section we use the original data collected by Jenks and Coulson [8] since their work sets a good benchmark for the visual quality of the classification. The data are rural population densities for minor subdivisions of Chase, Dickinson, Geary, Lyons, Marion, Morris and Wabaunsee countries in central Kansas (refer to Figure 5).

These data was classified by the 6 optimization strategies and the results are shown in Figure 6, Figure 7 and Figure 8.

The classification produced by the two approaches of MED is shown if Figure 9.

The maps of the optimization strategies here coincide with the result obtained in the recent comparison by Cromley [4]. For example, Table 1 shown the values obtained for each of the six optimization criteria by the optimization strategies and the two approaches with MED on the real data. Each optimization strategy naturally results in the one that minimizes the criteria is set to minimize. Note that MED is not much behind the optimum value in each case. It is also important is to point out that the maps for MED are much clearer of the overall situation in the map. Optimization strategies have a tendency to isolate values that are away from the majority of values and then, create groups that have to many elements, although interesting similarities occur between this elements. For example, the region with value 103.4 is placed by itself in several of the optimization strategies while details is lost in the range of values between 2 and 6, where many of the values are.

Algorithm	Metric Criteria					
	Var_{L_1}		Var_{L_2}		Var_{L_∞}	
	Total	MinMax	Total	MinMax	Total	MinMax
MED-U	71.8	18.0	102.9	34.8	9.0	2.4
MED-R	79.7	35.2	109.5	71.1	8.5	3.9
Total Var_{L_1}	60.0	14.8	99.5	50.6	11.0	4.05
MinMax Var_{L_1}	61.9	11.8	96.28	40.23	11.35	3.85
Total Var_{L_2}	72.1	28.6	88.33	28.5	9.1	3.6
MinMax Var_{L_2}	81.7	24.9	123	26.0	11.3	3.6
Total Var_{L_∞}	100.2	55.2	150.39	94.33	5.25	2.75
MinMax Var_{L_∞}	89.4	43.7	107.07	55.22	5.3	1.85

Table 1: Metric values obtained by the classification methods on the Jenks and Coulson real data set.

5 Final remarks

Aronoff [1] has postulated that the optimum of data quality is the minimum level of quality that will do the job. He has also indicated that the most cost-effective decision model is the simples model that does most with the least. Judging by these criteria, classification by MED as proposed here it is a clear winner. MED's computational cost is clearly superior to the optimization strategies, both in terms of computational time and space. In terms of the maps resulting form MED's classification, we observe here, with the benchmark data of Jenks and Coulson [8] that MED produces “good visual” maps since the large variations are apparent and no detail is lost in the lower values.

To our knowledge, the work reported here is the first attempt to extrapolate techniques based on information theory for classification of geographical data. This has also indicated the fundamental assumptions of our approach and further research issues are open by our presentation. Among these, the most immediate is the use of Maximum Likehood Estimation. Although this is perfectly acceptable for the classical statistician, the Bayesian statistician will argue that for each geographical region we have an “a priori” belief on the value of the probability, that is updated by new observations. The Bayesian model will be appropriate for GIS since for most cases, new maps are generated from older maps. Not all information in a GIS was obtained at the same time, and even if that is the case, the updates to the data in the GIS will perturb this. We expect to pursue this topic in the future.

References

[1] S. Aronoff. *Geographic Information Systems: A Management Perspective.* WDL Publications, Otttawa, Canada, third edition, 1993.

[2] D.K.Y Chiu, A.K.C. Wong, and B. Cheung. Information discovery through hieratchical maximum entropy discretization and synthesis. In G. Piatetsky-Shapiro and W.J. Frawley, editors, *Knowledge Discovery in Databases*, pages 125–140, Menlo Park, CA, 1993. AAAI, AAAI Press.

[3] M.R. Coulson. In the matter of class intervals for choropleth maps with particular reference to the works of George F. Jenks. *Cartographica*, 24:16–39, 1987.

[4] R.G. Cromley. A comparison of optimal classification strategies for choroplethic displays of spatially aggregated data. *International Journal of Geographical Information Systems*, 10(4):405–424, 1996.

[5] I.S. Evans. The selection of class intervals. *Transactions of the Institute of British Geographers*, 2:98–124, 1977.

[6] J.A. Hartigan. *Clustering Algorithms.* John Wiley, New York, 1975.

[7] A. Hitchinson. *Algorithmic Learning.* Clarendon Press, Oxford, UK, 1994. Graduate Texts in Computer Science.

[8] G.F. Jenks and M.R.C. Coulson. Class intervals for statistical maps. *International Yearbook of Cartography*, 3:119–134, 1963.

[9] M. Lascurain. *On Maximum Entropy Discretization and Its Application in Pattern Recognition.* PhD thesis, Dept. of System Design Engineering, University of Waterloo, Waterloo, N2L-3G1 Ontario, Canada, 1983.

[10] Cromley. R.G. and G.M. Campbell. Range constrained grouping of univariate data for choroplethic displays. In T. Waugh and R. Healey, editors, *Advances in GIS Research*, pages 341–355, London, UK, 1994. Taylor and Francis Publishers. Vol 2.

[11] C.E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27(3):20–23, 1948.

[12] C.E. Shannon and W. Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, Urbana, Ill., 1964.