CHOROPLETH MAPS: CLASSIFICATION REVISITED

Gennady Andrienko, Natalia Andrienko, and Alexandr Savinov

GMD - German National Research Center for Information Technology
Schloss Birlinghoven, Sankt-Augustin, 53754 Germany

URL: http://ais.gmd.de/KD/
e-mail: gennady.andrienko@gmd.de
fax: +49-2241-142072

Abstract

Classification was traditionally used as an instrument for producing choropleth maps. In our study we consider the role of the process of classification in exploratory data analysis. We developed a set of tools for classification that facilitate looking on data from various viewpoints and thereby investigate different aspects of the data. The tools include direct manipulation controls for specifying arbitrary class boundaries, graphs representing statistical distribution of attribute values, means for automatic classification, calculation of statistical quality of a classification, and various color schemes that can be applied to represent classes on a choropleth map. The classes can be dynamically propagated to other displays showing different characteristics of the same objects. This helps in exploring relationships between attributes. Additionally, relationships can be investigated by applying various data mining methods to the classes produced. All the instruments are used in a highly interactive and dynamic mode: results of each user's action lead to immediate update of all displays involved.

1. Introduction

Choropleth maps are extensively used for data presentation. Significant efforts have been invested into development of various methods of data classification for choropleth mapping. Historically, classification was used in order to minimize the number of colors needed for representing data values on printed maps. This goal together with the task of minimizing subjectivity of data representation lead to development of 3 widely used methods of data classification [1]:

1) classification into equal intervals;
2) classification with equal frequencies of objects in the classes;
3) statistically optimal classification.

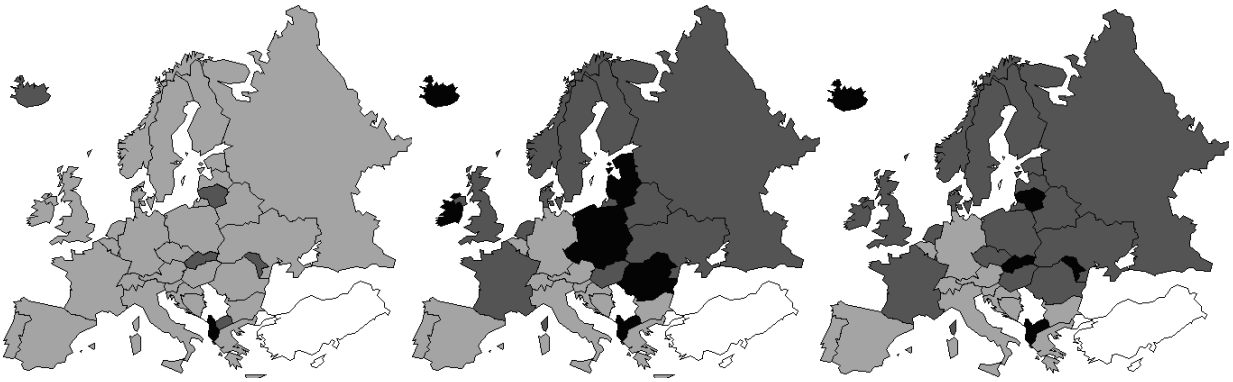


Figure 1. 3 variants of classification of European countries according to birth rates: equal interval classification, equal frequency classification, and statistically optimal classification.

It is well known in cartography that different selection of the number of classes and class breaks can radically change the information perceived from the map [2]. This can be seen, for example, from comparison of the choropleth maps shown in Figure 1. These maps result from three variants of classification of the European countries into 3 classes according to values of the attribute "birth rate", but it is rather hard to believe that they represent the same data. The effect of classification can be observed by comparing these maps to an unclassed choropleth map representing the same

attribute (Figure 2). So, classification is a powerful tool for cartographers allowing them to express different ideas they would like to communicate to map consumers [3].

Figure 2. An unclassed choropleth map represents values of birth rates by proportional degrees of darkness.

Nowadays, maps are more and more used for analysis and exploration of data rather than for data registry and communication. This poses new demands to design of thematic maps. In particular, an analyst cannot be satisfied with a static choropleth map with a fixed classification because such a map conveys just one of numerous spatial patterns possible (recall the maps in Figure 1). Among these patterns there are no “correct” and “wrong” ones: any of them can expose some important feature of the data. Therefore an analyst needs tools to experiment with different classification schemes and their parameters as well as to change arbitrary the boundaries between classes. Interactivity and fast feedback play a crucial role.

Recently a considerable progress has been made towards suiting classification methods and tools to the needs of data exploration. Egbert and Slocum [4] developed a system ExploreMap that is based on the use of classed choropleth maps. This system allows the user to specify the classes interactively and see the results on a map. Our system Descartes [5, 6] provides additionally statistics about available attributes for each class (minimum, maximum, average values etc.). Propagation of class colors to various statistical graphical displays is helpful in investigation of relationships between attributes (see, for example, [7]). This operation is an extension of "brushing", a widely recognized technique of exploratory data analysis. Relationships can also be studied by means of data mining methods applied to results of classification [8].

We propose an integrated framework for spatial data exploration on the basis of classification, which includes a variety of tools for classification and data exploration on the basis of classification:

1) measurement of statistical quality of a classification;
2) the traditional automatic classification methods: equal intervals, equal frequencies, and optimal classification;
3) direct manipulation facilities for interactive specification of breaks;
4) graphs that represent statistical distribution of attribute values: dot plot and cumulative frequency curve;
5) various color scales and interactive controls to manipulate them in order to make the maps more expressive;
6) dynamic brushing which links a classed choropleth map to additional graphical displays by using the same coloring of objects;
7) computation and visual presentation of summary statistics for the classes: number of objects, minimum, maximum, and average values of any selected attribute as well as medians and quartiles;
8) data mining methods that can be applied to classes produced.

2. Statistical quality of a classification: measurement and visualization

In the result of any classification the data loose their precision. The amount of precision lost (i.e. the statistical error introduced by the classification) can be measured. The idea of measurement of the statistical error of a classification was introduced into cartography by G.F. Jenks [9]. This was meant for use in forming classes that are internally homogeneous while assuring heterogeneity among classes [10]. The total error of a classification is calculated as a sum of internal errors for each individual class, $E = \sum E_k$. The internal class error E_k is calculated depending on the chosen measure of diversity. We introduced 3 measures of diversity named “mean measure”, “median measure”, and

"entropy measure". The first measure is calculated as a squared deviation from the class mean: $E_k = \sum (x_i - \overline{X}_k)^2$, where $\overline{X}_k$ is the class mean. The second measure is calculated as a sum of absolute deviations from the class median, $E_k = \sum | x_i - \tilde{X}_k |$, where $\tilde{X}_k$ is the class median. This measure is less affected by extremes in the tails of the distribution (outliers) than the first measure because the data in the tails have less influence on the median than on the mean. The third measure aggregates deviations from the mean using logarithm function conventional in information theoretic approaches, $E_k = \sum | x_i - \overline{X}_k | \log | x_i - \overline{X}_k + 1 |$. If the deviation is small the function increases approximately as the mean squared deviation (the first measure) while for large deviations it grows much slower, so that outliers make less contribution into the total error.

For each classification produced by the system or modified by the user we compute 2 indicators of the quality. The first represents the loss of precision in the result of the classification. The second is a ratio of the first value to the value for the statistically optimal classification with the same number of classes. Hence, the first indicator expresses how far the classification is from the original data set, and the second one - how far it is from the optimal classification. In order to find the optimal classification, we use the algorithm proposed by Fisher [10–12]. The algorithm has a special iterative organization when classifications obtained on earlier steps are efficiently used on the next steps.

Both quality indicators are presented to the user as numbers measured in the scale 0% to 100% as well as graphically (see Figure 3). The indicators are automatically updated after each change of the classes.

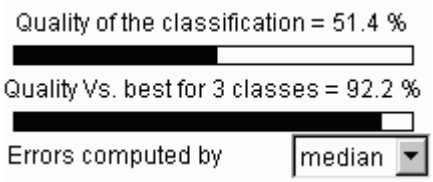


Figure 3. Representation of the statistical quality of a classification.

3. Interactive tools for classification

3.1. Specification of classes

The user has the following tools for defining classes:

1) Automatic classification methods: equal interval classification, equal frequency classification, and statistically optimal classification. Two ways of using these methods are possible: the user can either specify the desired number of classes or the desired quality. In the latter case the number of classes required will be automatically found by the system.
2) A text editing control which lists current values of the breaks. The user can edit the break values, remove breaks, and introduce new breaks within this control.
3) A compound slider [5, 6], which visually represents relative positions of the class breaks between the minimum and the maximum values of the attribute and displays the currently used color scheme (Figure 4). The control is supplied with a dot plot that represents statistical distribution of the attribute values.

Figure 4. Visual appearance of the compound slider. Three slider pointers (vertical lines with small triangles at both ends) divide the interval into 4 classes. In the dot plot above the slider the gray circles represent values of the attribute.

The compound slider is a direct manipulation control. The user can move any of slider pointers and in this way update the classification. When a pointer approaches one of its neighbors, it is removed (i.e. two consecutive classes are merged). When the user clicks in a free space between two slider pointers, a new class break is added that divides the respective class into two classes.

The compound slider has a special function: during the process of movement of a slider pointer only the objects belonging to the two classes affected by the corresponding boundary change are shown on the map by painting, while the remaining objects are shown in neutral gray color. This helps the user to concentrate better on the changes of the spatial patterns resulting from the movement of the boundary.

A statistics panel (Figure 5) shows to the user absolute and relative numbers of objects fitting in each class of the current classification. All figures are immediately updated after any change of the classes, in particular, during movement of a slider pointer. Optionally, the user may view additional statistical information in numeric or graphical form, for example, average value of death rate, sum of population numbers, and median life expectancy for each class. The system can design diagrams representing summaries for the requested attributes. For example, average age structure of the population in each class can be presented by pie charts.

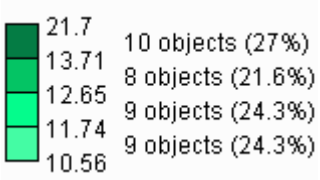


Figure 5. Statistics panel provides information about distribution of objects among the classes.

3.2. Frequency histogram and cumulative frequency curve

In classification of spatially referenced data an analyst needs to consider the data from two perspectives, statistical and spatial, and take into account the peculiarities of both the statistical and the spatial distributions of the data. This means that the analyst needs to pursue at least two concurrent goals. The first is to minimise variation of data within each class and to maximise differences between classes. The second goal is to divide the territory into the smallest possible number of coherent regions with low data variation within the regions. The analyst needs such tools that would allow her/him to balance between these goals in search of an acceptable compromise solution. Whereas the statistically optimal classification algorithm is suited to the first goal, it does not take into account the spatial aspect of the data and, hence, cannot help in approaching the second goal. Therefore so important is the “freehand” classification procedure described above: the user moves class breaks and immediately observes the changes of the patterns on a map. A visual representation of the statistical value distribution can help the analyst to meet also the statistical criteria in the course of interactive classification. The dot plot accompanying the compound slider could be suitable for these purposes, but overlapping of point symbols in it obscures understanding of the distribution.

Traditionally, statistical distribution is represented on frequency histograms. In a histogram numbers of objects with similar values of the attribute are represented by heights of bars. However, this graph has several disadvantages. First, the heights of the bar and the shape of the whole graph strongly depend on the selected granularity of the horizontal axis (Figure 6). Second, there are serious problems with estimation of the numbers of objects in classes. When the histogram shows frequency counts for pre-selected equal intervals that are different from the class intervals, it is necessary to add together the lengths of the bars fitting into one class interval. The estimation becomes impossible when class breaks fall between the interval boundaries on the histogram. When the histogram is built so that the bars correspond to the class intervals (that, in general, differ in lengths), the widths of the bars are proportional to the lengths of the intervals, and it is difficult to avoid estimation of the bar areas instead of the heights. Bars for longer intervals will always produce an impression of lager number of objects in the corresponding classes.

One more method for graphical representation of statistical distribution is the cumulative frequency curve, or ogive. In such a graph the horizontal axis represents the value range of an attribute. The vertical position of each point of the curve corresponds to the number of objects with values of the attribute being less than or equal to the value represented by the horizontal position of this point. Peculiarities of value distribution can be perceived from the shape of the ogive. Steep segments correspond to clusters of close values. The height of such a segment shows the number of the close values. Horizontal segments correspond to “natural breaks” in the sequence of values.

It is important that the cumulative frequency curve does not require prior classification, and the shape of the curve does not depend on the granularity of the axes. However, the graph *can* represent results of classification by means of additional graphical elements. In Descartes the horizontal axis of the graph is divided into segments to show classification intervals. The lengths of the segments are proportional to the length of the intervals. The segments are

painted in the colors of the classes. The positions of the class breaks are projected onto the curve, and the corresponding points of the curve are, in their turn, projected onto the vertical axis (Figure 7). The division of the vertical axis is also shown with the use of colored segmented bars. The lengths of the segments are proportional to the numbers of objects in the corresponding classes. With such a construction it becomes easy to compare the sizes of the classes. For example, the segmentation of the vertical axis in the middle panel of the Figure 7 clearly demonstrates that the corresponding classes have approximately equal numbers of objects.

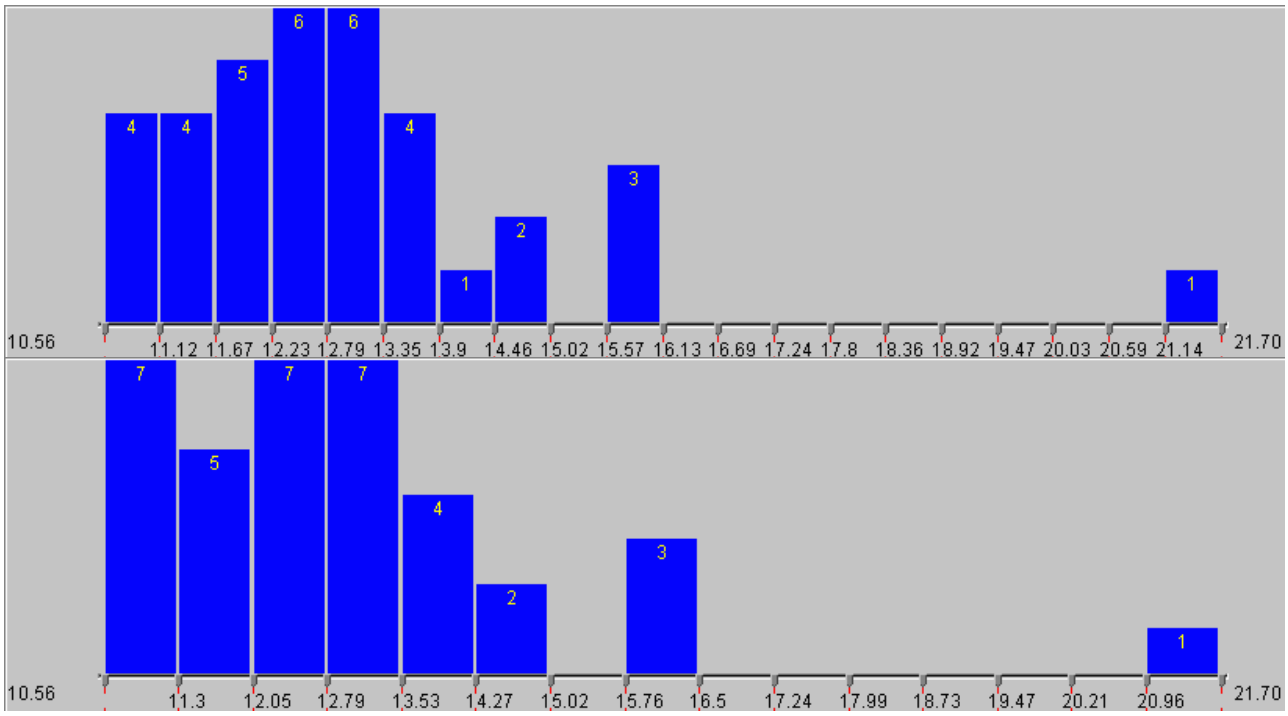


Figure 6. Two examples of the histogram plot built for the same data (distribution of birth rates in European countries) with granularity 0.15 and 0.20, respectively.

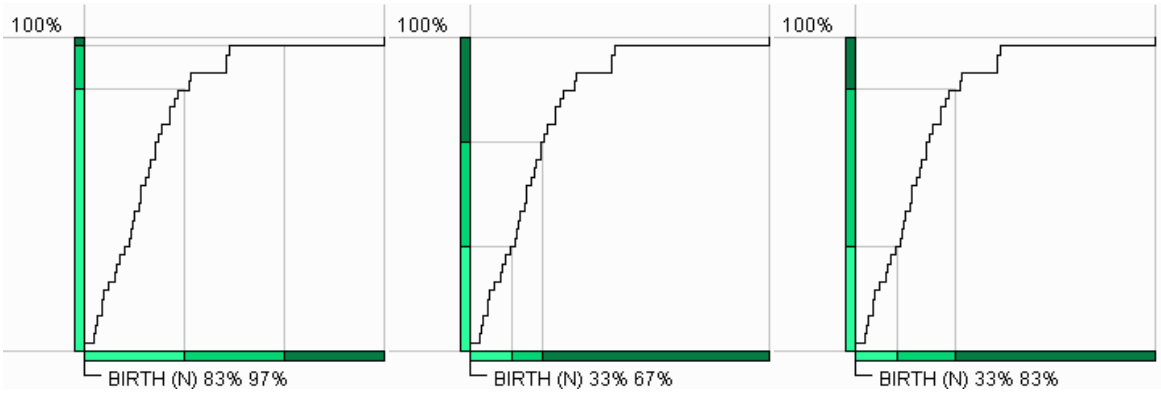


Figure 7. Cumulative frequency curve for the birth rates. The axes are divided into segments corresponding to current classes and painted into the colors of the classes. The three panels correspond to the variants of classification represented in Figure 1: equal intervals, equal frequencies, and optimal (median) classification.

3.3. Transformation of color scales

Results of a classification could be considered as a new attribute with an ordered scale of values. Respectively, the visualization of the classification should reflect this ordering in coloring of the classes. Appropriate for this purpose are single-hue color schemes with monotonously increasing or decreasing darkness that represents the ordering. Spectral or

multiple-color schemes with arbitrary degrees of darkness not corresponding to the ordering may confuse or mislead the user.

Interesting observations can be done using specially designed bi-directional color schemes [13]. Such schemes are used in cartography for emphasizing progressions outward from a critical midpoint of data (reference value). One color hue is used for values below the reference value and another hue for values above the reference value. The distances to the midpoint are represented by degrees of darkness. In exploratory data analysis, however, critical points in data are often previously unknown. Earlier we invented a 'dynamic comparison' technique for unclassed choropleth maps that can help to empirically reveal them [5]. The idea is that the user may interactively move the midpoint and immediately observe the changes of the spatial pattern on a map. In this way she/he can find such reference values that lead to meaningful spatial patterns. Recently we have adapted the visual comparison technique to classed choropleth maps. The counterpart of a reference value in a classed map is a reference class.

We developed a direct manipulation tool that allows the user to select and interactively change the reference class in the current classification. For this purpose the user needs to position the vertically oriented triangle (reference pointer) located below the slider used for classification. By moving the reference pointer the user can easily switch between different color scales: a single hue scale with increasing darkness, the same with decreasing darkness, and bi-directional scales with different positions of the midpoint. Figure 8 shows 6 variants of color scales for a classification into 4 classes and the corresponding positions of the reference pointer.

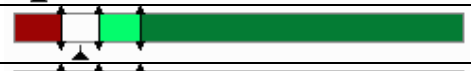
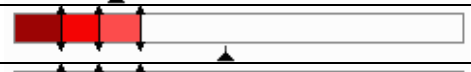
	Left end		Right end
1	light green		dark green
2	White		dark green
3	dark red		dark green
4	dark red		dark green
5	dark red		white
6	dark red		light red

Figure 8. Six different color scales for the same classification with 4 classes.

3.4. Dynamic brushing and paneling

Within the system several graphical displays may be simultaneously present on the screen. Besides maps, these may be dot plots, scatter plots, and parallel coordinate plots. The displays can represent different attributes of the same objects or the same attributes in different ways or combinations. Having a classification, it is possible to propagate class colors to other displays. As a result, points on a dot plot or on a scatter-plot and lines on a parallel coordinates plot will be painted in the colors of the classes the respective objects fit in. This feature, known as "brushing", is very useful for studying relationships between attributes (see Figure 9).

Another interesting variant of propagation of classes to graphical displays is paneling [7], that is, cloning of a display so that each copy of it shows one of the classes (Figure 10). In many respects paneling is superior to brushing. Usually it produces clearer views, in particular, when there are display areas where points or lines of objects belonging to different classes mix. Paneling can be applied even to plots that are not suitable for brushing, i.e. those using colors to represent other information than object classification. A panel can combine two or more classes, and the user can be given an opportunity to select which classes to view in each panel [14].

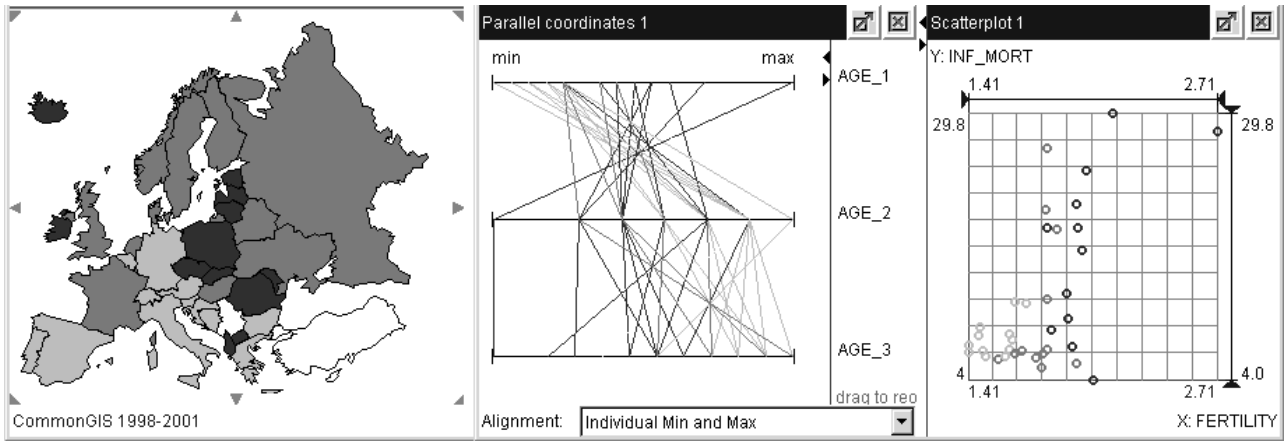


Figure 9. Propagation of classes to additional displays. The countries are classified into 3 equal frequency classes according to birth rates. Darker color corresponds to bigger values. The same colors are used to paint dots on the scatter plot on the parallel coordinate plot.

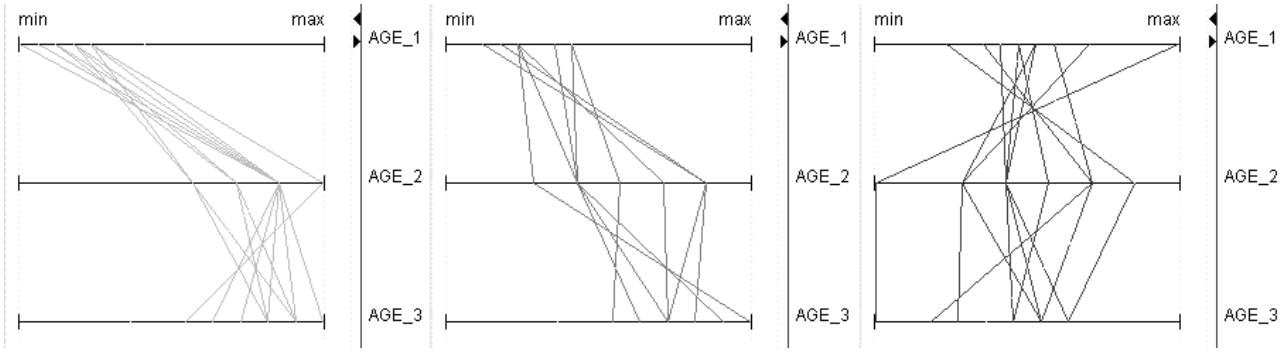


Figure 10. Paneling applied to the parallel coordinate plot from Figure 9. The classes are the same as in Figure 9.

3.5. Application of data mining methods to results of classification.

The link between our visualization system Descartes and a data mining system Kepler [15] makes it possible to apply different data mining algorithms to classes produced within our system. The architecture of the link is described in [8]. Just to give an example, we show in Figure 11 the result of application of a data mining algorithm called C4.5 [16] to the division of the European countries into 3 equal frequency classes according to birth rates.

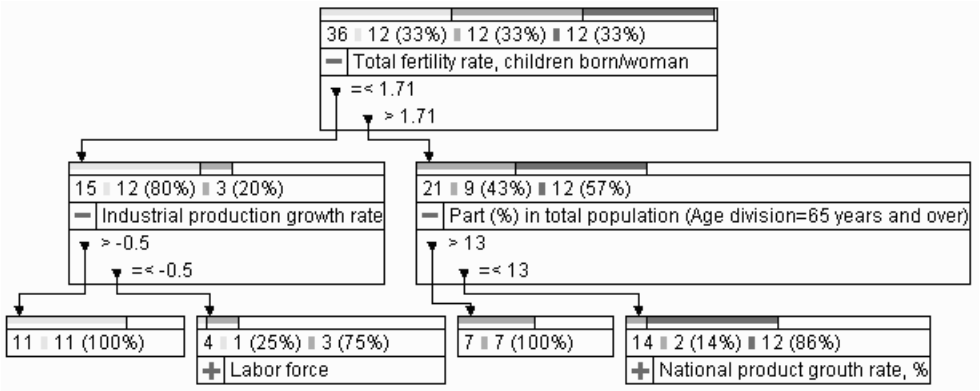


Figure 11. The decision tree resulting from application of the C4.5 data mining methods to data about the countries of Europe classified into 3 equal frequency classes according to birth rates.

The C4.5 algorithm tries to discriminate between the given classes (the origin of which is irrelevant) on the basis of the values of available attributes and to produce a decision tree that divides the whole set of objects into groups as close as possible to the given classes. Each tree node represents a step in set division based on values of one attribute. It is important to note that the C4.5 algorithm tries to find the most discriminative attributes. Due to this the decision tree may help to reveal important relationships in the data set under investigation. Thus, from the tree in Figure 11 one may see that all 12 countries with low birth rates are characterized by fertility rates below 1.71, and 11 of them have industrial production growth rates more than –0.5%. All the countries with high birth rates have less than 13 % of elderly people (65 years and over) in their population. More about the use of the C4.5 method in combination with interactive mapping facilities of Descartes can be read in [17].

4. Discussion

We have described here the array of complementary tools for classification implemented in our system. Very important are high interactivity of the tools, immediate reaction of all displays to any slight change of classes, and possibility of viewing data from multiple perspectives. This turns the procedure of classification into a powerful instrument of exploratory analysis of spatially referenced data.

We tested a part of our tools for classification during a validation study within the CommonGIS project (Esprit project 28983, 1998 - 2001). We did not include in the study the use of cumulative curves and the link to data mining. The study was performed in CNIG (Portugal) in March 2001. After a short acquainting lection and demonstration of the tools the subjects were well able to use them for producing classifications with requested properties. At the same time certain user interface problems have been revealed. Thus, the large number and variety of available controls impede their efficient use. Besides, some people encountered difficulties in manipulation of color scales. Now we work on solution of these problems. In particular, we strive to make the system configurable and customizable to users' needs and skills.

Implementation of the system is done in Java. The system will be demonstrated at the conference. Additionally, it can be accessed through the Internet.

Acknowledgements

The work was partly supported by the European Commission in projects "CommonGIS – Common Access to Geographically Referenced Data" (Esprit project 28983, 1998 - 2001) and "SPIN ! - Spatial Mining for Data of Public Interest" (IST Program, project No. IST-1999-10536, 1999 - 2002)

References

1. Robinson, A.H., Morrison, J.L., Muehrcke, P.C., Jon Kimerling, A., and Guptil, S.C. (1995*) Elements of Cartography.* New York: Wiley.
2. MacEachren, A.M. (1994). *Some Truth with Maps: A Primer on Symbolization and Design.* Association of American Geographers, Washington
3. Yamahira, T., Kasahara, Y., and Tsurutani, T. (1985) How map designers can represent their ideas in thematic maps. *The Visual Computer*, **1** (1), pp.174-184
4. Egbert, S.L. and Slocum, T.A. (1992) EXPLOREMAP: an exploration system for choropleth maps. *Annals of the Association of American Geographers*, **82**, pp.275-288.
5. Andrienko, G. and Andrienko, N. (1999a) Interactive Maps for Visual Data Exploration. *International Journal Geographical Information Science*, 1999, **13** (4), pp.355-374
6. Andrienko, G. and Andrienko, N. (1998) Dynamic Categorization for Visual Study of Spatial Information. *Programming and Computer Software*, 1998, **24** (3), pp.108-115
7. Wilkinson, L. (1999) *The Grammar of Graphics.* New York: Springer.
8. Andrienko, G. and Andrienko, N. (1999b) Knowledge-Based Visualization to Support Spatial Data Mining. In Hand, D.J., Kok, J.N., and Berthold, M.R. (Eds.*) Advances in Intelligent Data Analysis 3rd International Symposium, IDA-99*, Amsterdam, The Netherlands, August 9-11, 1999, Proceedings. Lecture Notes in Computer Science, vol. **1642**. Berlin: Springer-verlag, pp.149-160
9. Jenks, G.F. (1977) *Optimal data classification for choropleth maps*: Occasional Paper No. 2, Dept. Geography, Univ. Kansas, 24 p.
10. Fisher, W.D. (1958) On grouping for maximum homogeneity: *Jour. Am. Stat. Assoc*. v. **53**, no. 284, pp.789–798.

11. Fisher, W.D. (1969) *Clustering and aggregation in economics*: John Hopkins Press, Baltimore, Maryland, 195 p.
12. Hartigan, J.A. (1975) *Clustering algorithms*: John Willey & Sons, New York, 351 p.
13. Brewer, C.A. (1994) Colour use guidelines for mapping and visualisation. In *Visualisation in Modern Cartography* (New York: Elsevier Science Inc.), 123-147.
14. Brunsdon, C. The Comap: Investigating Geographical Pattern via Conditional Spatial Distributions. Proceedings GISRUK'2000, pp.97-103
15. Wrobel, S., Wettschereck, D., Sommer, E., and Emde, W. (1996) "Extensibility in Data Mining Systems", *Proceedings of KDD'96 2nd International Conference on Knowledge Discovery and Data Mining*, AAAI Press, pp.214-219
16. Quinlan, J.R. (1993). *C4.5 Programs for Machine Learning*. Morgan Kaufmann, 1993
17. Andrienko, G. and Andrienko, N. (1999c) "Data Mining with C4.5 and Cartographic Visualization", in N.W.Paton and T.Griffiths (eds.) User Interfaces to Data Intensive Systems, IEEE Computer Society Los Alamitos, CA, pp. 162-165.