PERGAMON

Computers, Environment and Urban Systems
25 (2001) 5–15

Computers, Environment and Urban Systems

www.elsevier.com/locate/compenvurbsys

Exploring spatial data with dominant attribute map and parallel coordinates

G. Andrienko *, N. Andrienko

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

Abstract

This paper suggests a technique and a corresponding software tool for exploratory analysis of multivariate numeric data associated with area geographical objects such as units of administrative division of a territory. More specifically, the paper considers analysis of several comparable attributes, i.e. those measured in the same units and semantically related. Examples of comparable attributes can be provided by attributes representing proportions in total (such as division of population by age group or land use statistics) or rates (such as birth and death rates). The technique suggested is based on the use of a parallel coordinate plot dynamically linked to an interactive map by means of simultaneous highlighting of corresponding objects. The map represents by painting classification of the geographical objects according to the dominant attribute: the colour of an object corresponds to the dominant attribute while the degree of darkness shows the strength of the dominance. The tool described enables various transformations of the parallel coordinate plot. The transformations help the user to consider data from different perspectives. Any transformation is immediately reflected by the map. The map and the parallel coordinate plot mutually facilitate the interpretation and enhance analysis-supporting capabilities of each other.

Keywords: Exploratory data analysis; Interactive maps; Parallel co-ordinates; Linked displays

1. Introduction

One of the mapping techniques that proves most useful for exploratory analysis is choropleth map, i.e. painting regions in colours according to values of attributes. The major advantage of this technique is that it allows the map reader to grasp the picture as a whole (Gestalt) and gain an overall view on the data (Bertin, 1967), i.e. it helps in revealing and understanding of global tendencies in data. Most often the

* Corresponding author. Tel.: +49-2241-142329; fax: +49-2241-142072.
E-mail addresses: gennady.andrienko@gmd.de, http://borneo.gmd.de/and/ (G. Andrienko),.

0198-9715/01/$ - see front matter
PII: S0198-9715(00)00037-5

technique of choropleth map is used to represent values of a single attribute. In traditional "paper" maps values of the attribute are typically classified into ranges (usually from 3 to 7), which are then matched to different degrees of darkness or sometimes colour hues. More suitable for exploration purposes is direct mapping from values to proportional degrees of darkness (continuous scale). The human eye could successfully catch peculiarities of spatial distribution on such a map and detect clusters of similar objects (Lewandowsky & Behrens, 1995).

Less usual and more difficult to interpret are choropleth maps representing two attributes. The description of the method can be found in (Brewer, 1994). Its interactive software realisation was done in the system DESCARTES (Andrienko & Andrienko, 1999a). The method has strong limitations. It always requires classification of value ranges of the attributes, and the number of classes should be small (from 2×2 up to 5×5) to ensure that the map is interpretable.

In certain cases colour is used for the representation of three attributes, specifically, when the attributes are parts of some meaningful whole (for example, gross domestic product in industry, agriculture and services). In this case the so-called three-variable balance colour scheme is applied: three different hues associated to individual attributes are mixed in varying proportions (Brewer, 1994). This kind of map also requires classification and is perceivable only when the number of differing colours resulting from the mixing is rather small.

In this paper we introduce a painting-based data visualisation method, which is applicable to several (more than two) comparable attributes. The method consists of ascribing an object to a class according to the value of the dominant attribute. There are several ways of determining the dominant attribute. In the simplest case the attribute with the largest value is dominant. The other approaches involve prior normalisation of attribute values.

The dominant attribute mapping method is intended to support the following exploratory activities:

1. overall view on spatial co-distribution of attribute values;
2. finding spatial clusters of objects similar to each other in terms of the considered attributes; and
3. detecting objects with anomalies or disproportion among the values of the attributes.

Exploratory analysis of spatial data can be greatly facilitated when visualisation in geographic space is dynamically linked to presentation of data in attribute space. Monmonier (1989) suggested using the brushing technique for linking maps and statistical graphics by simultaneous highlighting of corresponding objects. Most often linking between maps and dot plots or scatter plots is considered (Andrienko & Andrienko, 1999a; Buja, Cock & Swayne, 1986; Dykes, 1997). Simultaneous representation of more than two variables can be done using parallel coordinates plots (PCPs; Inselberg, 1985). This kind of graphic is very useful for visual data exploration and data mining (Avidan & Avidan, 1999; Inselberg, 1998). MacEachren, Wachowicz, Edsall, Haug and Masters (1999) describe an implementation of a

link between PCP and one-variable map via brushing. We suggest some further modifications of PCP and a link of PCP to dominant attribute maps. In this combination, PCP, in addition to its traditional usage, also helps the user to understand and interpret the map.

The dominant attribute mapping method and its link to a PCP display are realised within DESCARTES, a knowledge-based system designed to support visual exploration of data with the use of maps (Andrienko & Andrienko, 1999a). An online demonstrator of the system, including all the examples cited in the paper, is available on the Internet (http://allanon.gmd.de/and/java/iris/).

Section 2 gives an overview of the data set used in the examples as well as of the parallel coordinates technique and the link between a map and a parallel coordinates display. Then we proceed with the description of proposed transformations of PCP (Section 3), and an explanation of the dominant attribute representation method (Section 4), followed by conclusions and directions of further development. An example of the application of the method to another data set is given in the Appendix A.

2. Population statistics of Portugal and ways of its visualisation

Within the CommonGIS project[1] (Andrienko & Andrienko, 1999b) the system DESCARTES is being applied to the data of the Census of Portugal taken in 1991. The data has been provided by CNIG — Portuguese National Centre for Geoinformation (http://www.cnig.pt/). In this paper, we particularly consider data about division of population according to age and employment in different branches of the economy. The data refers to 275 municipalities of the NUTS 4 level of administrative division.

Fig. 1 shows spatial distributions of percentages of population in four age groups: children (0–14 years), young people (15–24 years), middle age (25–64 years), and old

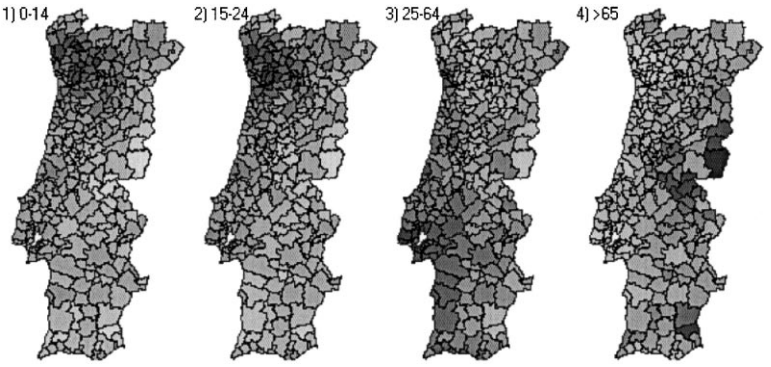


Fig. 1. Spatial distributions of percentages of population in different age groups.

[1] Esprit project No 28983, November 1998–April 2001, cf. http://commongis.jrc.it/.

(more than 65 years). The data is represented using the "small multiples" method (Tufte, 1983). Each attribute is shown on a separate unclassified choropleth map. This presentation method is good for surveying spatial distribution of values of each attribute and for comparing distributions. Thus, in our example, one can see that the biggest proportions of children and young people are in the North, middle-age population prevails in the West, and old people in the East.

In Fig. 2 the same data are represented in a PCP. The axes, in the top-down order, correspond to the age groups 0–14, 15–24, 25–64, and more than 65 years. Each line (trajectory) in the plot characterises some spatial object, i.e. municipality. The line connects the points on the axes that represent the values of the attributes associated with this object. The presentation by PCP gives an idea about relationships among the attributes. Thus, the prevailing parallel lines between two axes may indicate positive correlation while diagonal lines may mean negative correlation. PCP also enables the comparison of objects and detecting groups of objects having similar values for some of the attributes.

PCP is linked to maps in two complementary ways:

1. Mouse-over linking. When the mouse cursor points at some object in the map, the corresponding trajectory is highlighted in the PCP. When the mouse approaches some line in the PCP, the relevant object is marked on the map.
2. Durable selection. It is possible to select some object(s) to be marked both on the map and in the PCP, irrespective of further mouse movements. This is useful for comparison of two or more objects as well as for examining characteristics of a cluster of neighbouring objects selected in the map.

For example, marked in Fig. 2 are the Portuguese municipalities, which have small ageing populations. The selection has been done by clicking on the left of the

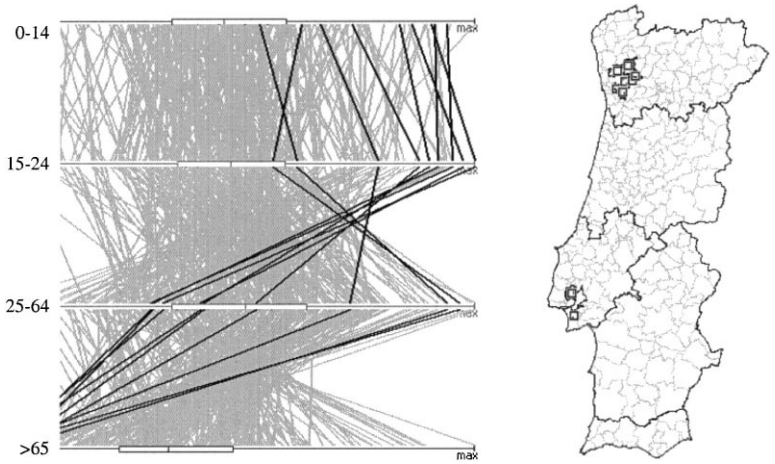


Fig. 2. A PCP display linked to a map. Highlighted (shown in black) are lines representing objects with the smallest percentages of old population.

bottom axis of the plot. One can see the distribution of values of the other attributes in the PCP, and geographical positions of the selected municipalities in the map (they are marked with black squares).

3. Transformation of parallel coordinates

Various modifications of PCP have been proposed lately (for example, Hoffman & Grinstein, 1999). Thus, some analytical tasks could be supported by radial or circular location of axes. We suggest some new variants of PCP that are particularly suitable for "comparable" numeric attributes. Attributes are comparable if they are measured in the same units and semantically related so that calculating numeric differences between their values makes sense. For example, the attributes "Birth rate (births/1000 population)" and "Death rate (deaths/1000 population)" are comparable. A special case of comparability is given when attributes represent parts of a meaningful whole, like proportions of different age groups in total population.

Since the attributes we deal with are comparable, we can bring the attribute axes to a common scale (see Fig. 3 left). Now the left end of each axis corresponds to the minimum value among all the attributes, and the right end to the maximum value among all the attributes. The positions on all axes having the same distance from the respective left ends represent one and the same number. Such a presentation especially supports overall comparison of variations of values of the attributes, as well as, comparison of individual values of the attributes associated with a selected object. Thus, in Fig. 3 one can see that the part of the middle-age population is bigger than the part of any other age group for all municipalities. It can be seen that percentages of age groups 0–14 and 15–24 years have similar ranges of variation, while the range of percentages of old people is about two times wider.

Another useful variant of scaling the axes is vertical alignment of the medians and quartiles of the attributes (Fig. 3 right). The rest of the values are linearly

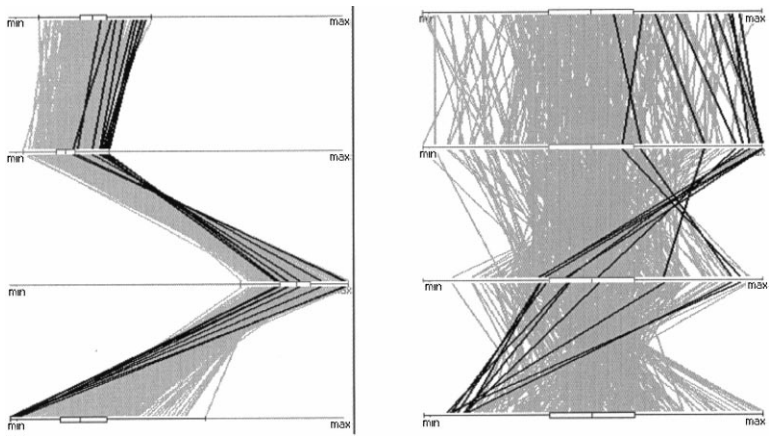


Fig. 3. Transformation of a PCP. Selected (shown in black) are the same municipalities as in Fig. 2.

interpolated along the scale. A similar picture could be obtained using the mean values and the standard deviations. This presentation is based on normalisation of values of the attributes. In fact, normalisation is also used in the traditional PCP when minimum and maximum values of attributes are aligned. This is equivalent to the division of all values of each attribute by the difference between its maximum and minimum values. The drawback of such a transformation is that it is very sensitive to outliers, i.e. values being significantly larger or smaller than the bulk of the values. Normalisation on the basis of median and quartiles ignores the lower and the upper one-fourths of the values and is therefore insensitive to outliers. Still, normalisation through minimum and maximum values is also useful in exploratory analysis. In particular, it is especially suitable for detecting outliers and objects with anomalies in proportions of values.

In order to facilitate various analytical activities, DESCARTES gives an opportunity to switch dynamically between different variants of scaling for parallel coordinates: without normalisation (common scale), normalised by minimum and maximum values, and normalised by median and quartiles.

To support investigation of data also in the geographical space, DESCARTES accompanies a PCP with a map presenting classification of the spatial objects according to the dominant attribute.

4. Mapping of dominant attributes

When the trajectory of some object in the attribute space is shown in a PCP, we can visually identify which attribute has the smallest or the biggest value, as well as view the difference between the extremes and the values of the rest of the attributes. In the PCP the rightmost vertex of the trajectory line corresponds to the attribute with the biggest value, the leftmost — to that with the smallest value. The shortest horizontal distance between the rightmost (leftmost) vertex and the rest of the vertices shows the size of the difference.

Let us consider, for example, the PCP with the common scale for all the attributes (Fig. 3 left). It is clear that in all municipalities the middle-age population dominates the other age groups: the value range of this attribute is located to the right of the value ranges of all the other attributes. However, the difference between the percentage of middle-age population and the maximum value among the other age groups significantly varies among the municipalities. We can visualise the degree of this difference (degree of dominance) in a choropleth map. Fig. 4 presents such a map linked to a PCP. Selected in both the displays are five municipalities having small differences between percentages of middle-age population and percentages of people in the next largest age group (i.e. old population in this case).

The same principle for selection of dominant attribute and calculation of the degree of dominance can be also applied to previously normalised values of attributes. Fig. 5 shows the results of classification of municipalities according to the dominant attribute, on the basis of minimum–maximum normalisation. DESCARTES represents such a classification by painting objects on the map in colours assigned to

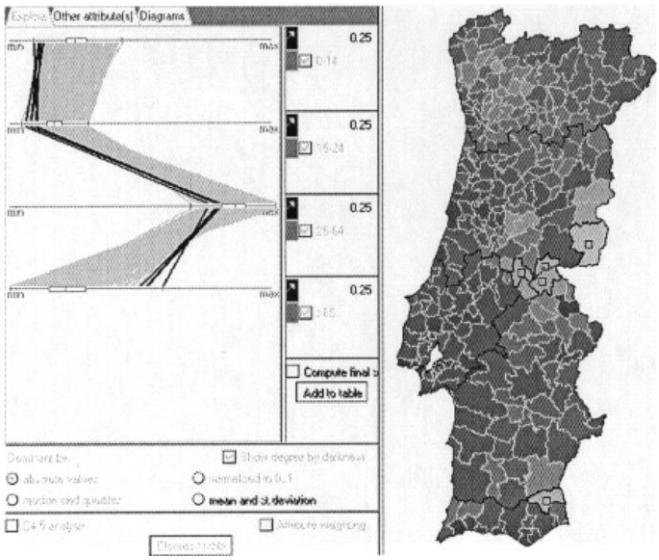


Fig. 4. Degree of dominance of middle-age population over the other age groups is shown by degree of darkness on the choropleth map. The map is dynamically linked to the PCP on the left.

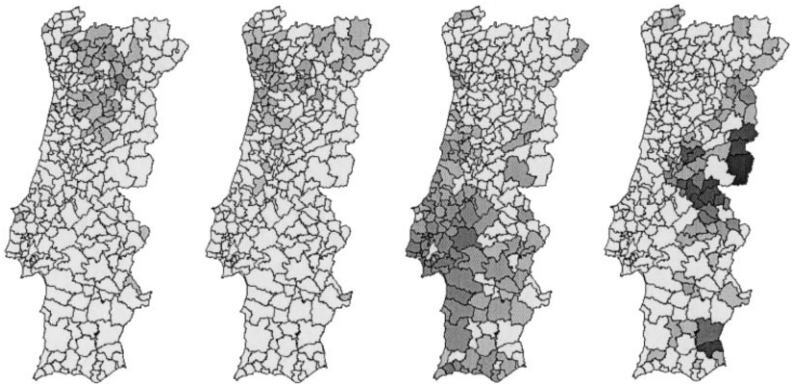


Fig. 5. Results of classification of municipalities according to the dominant attribute with prior minimum–maximum normalisation of the attributes. In order left to right: 0–14 years, 15–24 years, 25–64 years, and 65 years and over. Each map represents by shading the strength of the dominance.

the attributes. The colour hue of an object indicates which attribute is dominant while the degree of darkness shows the degree of dominance, or, in other words, the degree of membership of the object in the corresponding class. In a black-and-white reproduction such a multi-coloured map would be completely illegible. Therefore, in Fig. 5 the distribution of objects of each class is shown on a separate map. In a similar manner Fig. 6 demonstrates the results of classification by dominance on the basis of median-quartiles normalisation.

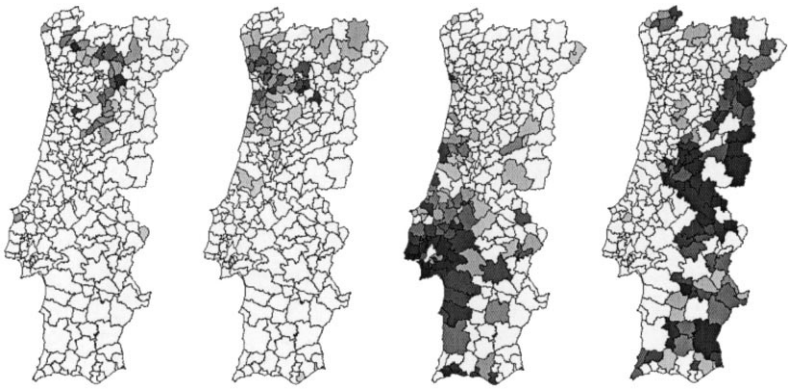


Fig. 6. Dominant attribute maps built on the basis of median and quartiles normalisation.

In both Figs. 5 and 6 we can observe spatial patterns similar to those visible in Fig. 1. It should be borne in mind, however, that the sequence of four maps in Figs. 5 or 6 has been artificially made for printing purposes. Actually it stands for a single map display in DESCARTES that shows all classes simultaneously (by painting in different colour hues) and thus gives an overall view on co-distribution of all the attributes. This display, in combination with the parallel coordinates plot, helps to detect spatial clusters of similar objects, in terms of the attributes considered. It is also easy to locate anomalies in proportions among the values of the attributes. These are objects painted in very dark shades in any of the maps.

Generally, the proposed variants of determining the dominant attribute (without normalisation, with minimum–maximum normalisation, or with median and quartiles normalisation) are complementary, and it is reasonable to view the results produced by all of them in the course of data exploration.

In the above example we used relative indices, specifically the percentages of different parts of a whole. To investigate the applicability of the method, we also applied it to absolute data: numbers of people occupied in agriculture, industry, and in services. In Fig. 7 spatial distributions of values of the attributes are presented using the "small multiples" method. A few very high values occur around Lisbon and on the North make two of the three maps not particularly expressive.

We considered all the three variants of determining the dominant attribute, and they resulted in similar classifications. Fig. 8 presents the results of classification by dominance among the non-normalised values (i.e. with common attribute scale). This classification vividly separates industrial, agricultural, and tourist areas of the country. Note that this division was not visible in the "small multiples" presentation.

5. Conclusions

In the paper we described a method of visualisation of multivariate data by means of a dominant attribute classification map linked to a parallel coordinate plot. This

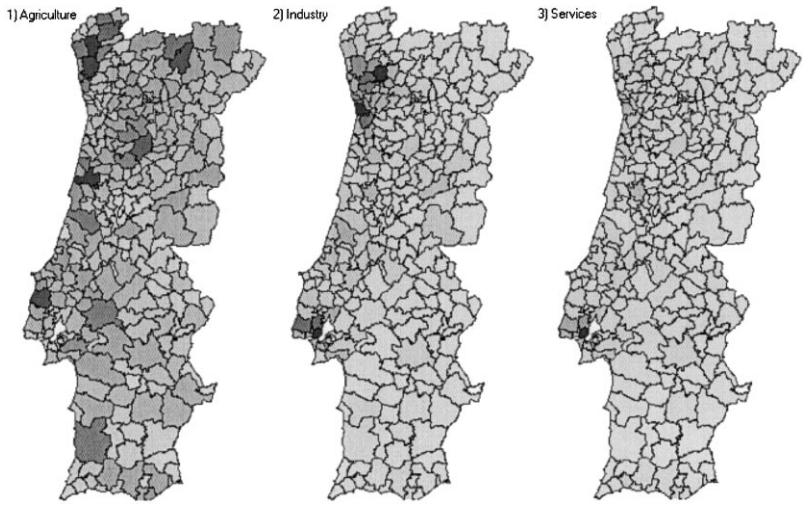


Fig. 7. Choropleth maps showing numbers of employees in agriculture, industry and services.

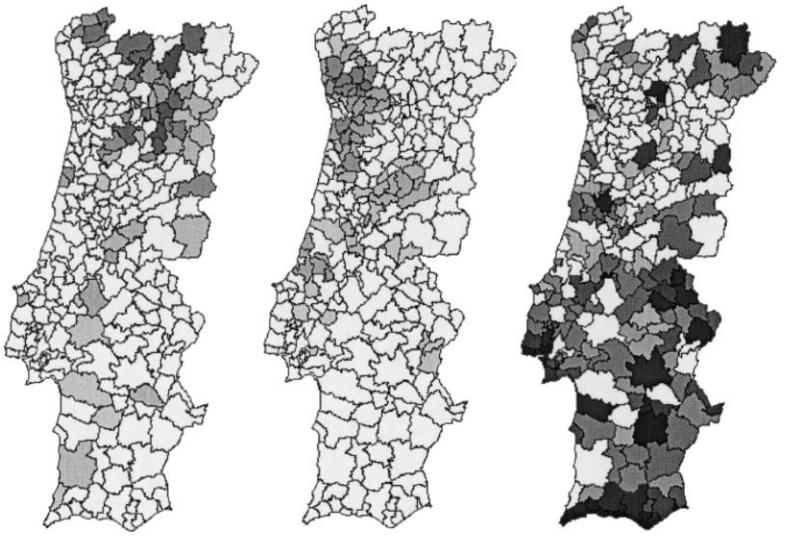


Fig. 8. Classification according to the dominant attribute made without prior normalisation. Agricultural, industrial and tourist areas are clearly separated.

method was successfully applied to different data sets and showed its usefulness for data exploration. Dominant attribute mapping offers a kind of data summary that may be helpful in certain analytical tasks such as getting a general idea about spatial co-distribution of several attributes, detecting clusters of objects with similar properties, and finding outstanding objects significantly differing from others.

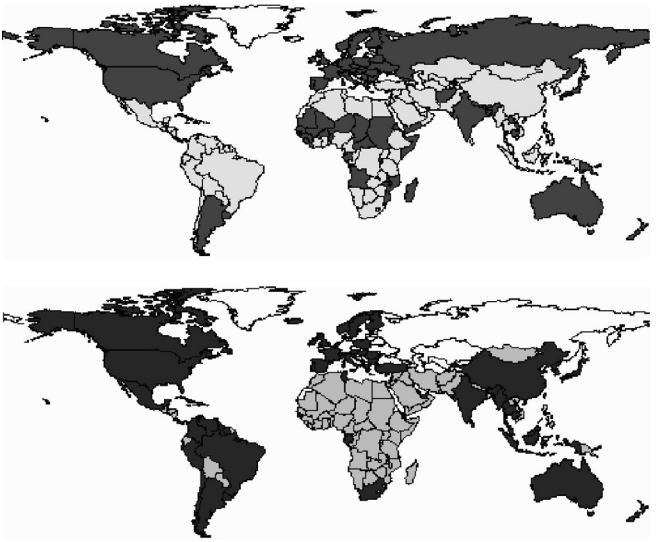
In the future we plan further development of the method. Thus, in addition to classification by the largest value, the system might propose to classify the objects according to the attribute with the smallest value. This will help to detect negative deviations from the bulk of the values. Another direction of development is integration with data mining methods. For example, the classification tree derivation algorithm may relate the produced classifications to other attributes (Andrienko & Andrienko, 1999c, d). The suggested method can be also integrated in a multi-criteria spatial decision support system. It may help a decision maker to analyse the set of feasible solutions in the criteria space and in the geographic space.

Acknowledgements

We are grateful to Dr. Voss (GMD), Gatalsky (GMD), Dr. Gitis (IPPI RAS), Prof. Jankowskij (University of Idaho) and Prof. MacEachren (Penn State University) for constructive discussions and suggestions concerning the proposed method.

Appendix A: Two maps of the World

The first map shows in light grey countries with domination of birth rates over death rates and in dark grey countries where death rates dominate over birth rates. In the second map domination of young population is shown by light pointing, and domination of middle-age people by dark. Both classifications are based on median-quartiles normalisation. The data originate from the ESRI world statistics set.



References

Andrienko, G., & Andrienko, N. (1999a). Interactive maps for visual data exploration. *International Journal Geographical Information Science, 13*(4), 355–374.

Andrienko, G., & Andrienko, N. (1999b). Making a GIS intelligent: CommonGIS project view. In *Proceedings of AGILE'99 Conference* (Rome, 15–17 April, 1999), AGILE, pp. 19–24.

Andrienko, G., & Andrienko, N. (1999c). Knowledge-based visualization to support spatial data mining. In D. J. Hand, J. N. Kok, & M. R. Berthold, *Advances in intelligent data analysis, lecture notes in Computer Science, vol. 1642* (pp. 149–160). Berlin: Springer-verlag.

Andrienko, G., & Andrienko, N. (1999d). Data mining with C4.5 and cartographic visualization. In N. W. Paton, & T. Griffiths, *User interfaces to data intensive systems* (pp. 162–165). Los Alamitos: IEEE Computer Society.

Avidan, T., & Avidan, S. (1999). ParallAX - A data mining tool based on parallel coordinates. *Computational Statistics, 14*(1), 79–90.

Bertin, J. (1967). *Semiology of Graphics. Diagrams, networks, maps*. Madison: The University of Wisconsin Press.

Brewer, C. A. (1994). Colour use guidelines for mapping and visualisation. In *Visualisation in modern cartography* (pp. 123–147). New York: Elsevier Science.

Buja, A., Cook, D., & Swayne, D. F. (1996). Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics, 5*, 78–99.

Dykes, J. A. (1997). Exploring spatial data representation with dynamic graphics. *Computers & Geosciences, 23*(4), 345–370.

Hoffmann, P., & Grinstein, G. (1999). Dimensional anchors: a graphic primitive for multidimensional multivariate information. David S. Ebert, & Christopher D. Shaw, *Workshop on New Paradigms in Information Visualization and Manipulation* (pp. 11–43). Kansas City, Missouri, USA.

Inselberg, A. (1985). The plane with parallel coordinates. *Visual Computer, 1*(1), 69–97.

Inselberg, A. (1998). Visual data mining with parallel coordinates. *Computational Statistics, 1*, 47–64.

Lewandowsky, S., & Behrens, J. T. (1996). Visual detection of clusters in statistical maps. In *Proceedings of the 1995 Meeting of the American Statistical Association* (pp. 8–17). Alexandria, VA: American Statistical Association.

MacEachren, A. M., Wachowicz, M., Edsall, R., Haug, D., & Masters, R. (1999). Constructing knowledge from multivariate spatiotemporal data: integrating geographic visualization with knowledge discovery in database methods. *International Journal Geographical Information Science, 13*(4), 311–334.

Monmonier, M. (1989). Geographic brushing: enhancing exploratory analysis of the scatterplot matrix. *Cartographycal Analysis, 21*, 81–84.

Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire: Graphics Press.