Interactive Cumulative Curves for Exploratory Classification Maps

Gennady Andrienko and Natalia Andrienko

Fraunhofer Institute AIS
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
Tel +49-2241-142486, -142329
Fax +49-2241-142072
e-mail gennady.andrienko@ais.fraunhofer.de
URL http://www.ais.fraunhofer.de/and/

Abstract. The paper addresses the problem of classification of geographical objects according to values of a numeric attribute. The authors are particularly interested in the role of classification as an instrument for exploratory analysis of spatially referenced data. Thus, through interactive classification one can investigate spatial distribution of attribute values and reveal spatial patterns. In this activity an analyst pursues at least two concurrent and often conflicting 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. A cumulative frequency curve may be used in the classification process in combination with a map to support the search of compromise solutions. Furthermore, the idea of the cumulative frequency curve can be extended to represent not only frequency but in general any quantitative characteristics of the objects being classified: cumulative area, cumulative population (in a demographic application) etc. With such curves one can investigate relationships between the attribute used for classification and other attributes. We propose an interactive tool for classification based on the use of the cumulative frequency curve and generalised cumulative curves.

1. Classification as an instrument for exploratory analysis

Representation of values of a numeric attribute referring to geographical objects, in particular, areas of territory division, is often done in cartography using the technique of classification. According to this technique, the value range of the attribute is divided into intervals. Then different colours are chosen to represent values from each of the intervals on a map.

Historically, classification was indispensable due to technical limitations involved in production of paper maps as well as in display of maps on early graphical computer screens (Robinson 1995). With appearance of modern computer hardware these limitations were eliminated, and it became possible to produce unclassed maps. In such maps numeric values are encoded by proportional degrees of darkness.

The merits and flaws of classed and unclassed maps have long been a topic of hot debates in cartography. It is not our intention to recite here this discussion or to take either side in it. Although classed maps are in the focus of this paper, this does not mean that we regard them as superior to unclassed maps. Our interest is the use of maps in exploratory analysis of spatially referenced data. In such analysis each type of maps plays its own role, and, hence, there is no question whether to select a classed or an unclassed map. An analyst should use both because these are complementary instruments of analysis.

Our opinion can be substantiated as follows. The goal of exploratory analysis is to gain understanding of given data, that is, to derive a short (compressed) description of their essential characteristics. Thus, according to Bertin, understanding is "discovering combinational elements which are less numerous than the initial elements yet capable of describing all the information in a simpler form" (Bertin 1965/1983, p.166). With regard to spatial distribution of values of a numeric attribute, an analyst initially has one value per each spatial object. The description of this data set could be substantially simplified if there were a directional trend in value distribution, for example, increase of values from the north to the south or from the centre of the territory to its periphery. Alternatively, a shorter description could be derived if the territory could be divided into possibly smaller number of coherent regions with low variation of attribute values within the regions. This technique is known as regionalisation[1]. Unclassed maps are better suited for detecting trends because they do not hide differences. Classification discards differences between values within a class interval and gives the corresponding objects similar appearance on the map. When these objects are geographical neighbours, they tend to be visually associated into clusters. This property makes classed maps well suitable for regionalisation. Which of the two ways to simplification occurs to be possible or more effective in each specific case, depends on the data and not on the preferences of the analyst. Therefore it is necessary to have both an unclassed choropleth map and a classification tool in order to investigate properly data with previously unknown characteristics.

It is clear, however, that a single static classed map cannot appropriately support regionalisation. It is well known in cartography that different selection of the number of classes and class breaks may radically change the spatial pattern perceived from the map (MacEachren 1994, Slocum 1999). There is no universal recipe of how to get an "ideal" classification with understandable class breaks, on the one hand, and interpretable coherent regions, on the other hand. Therefore when we say that classification may be used as an instrument of data analysis, we mean not a classed map by itself but an interactive tool that allows the analyst to change the classes and to observe immediately the effect on the map.

The exploratory value of classification was recognised in cartography only relatively recently. Initially classification was regarded as a tool for conveying specific messages from the map author to map consumers. Thus, the paper (Yamahira, Kasahara, and Tsurutani 1985) considers various possible intentions of the map designer and demonstrates how they can be fulfilled through application of different classification methods and selection of the number of classes.

In early nineties Egbert and Slocum developed a software system called ExploreMap intended to support exploration of data with the use of classed choropleth maps (Egbert and Slocum 1992). The most important feature of the system was a possibility to interactively change the classes. Another implementation of this function based on direct manipulation techniques is the "dynamic classification" tool incorporated in the system CommonGIS (Andrienko and Andrienko 1998, 1999). Exploration on the basis of classification is additionally supported in CommonGIS by the function of computing statistics for the classes: the range of variation and the average, median and quartile values of any selected attribute for each class.

In this paper we describe a recently developed extension of the dynamic classification tool that exploits the properties of the cumulative frequency curve and generalised cumulative curves. In

[1] Regionalisation can be (and often is) applied not only to values of a numeric attribute but to any spatially distributed data.

the next section we define the relevant notions and explain the use of cumulative curves in classification and data exploration.

2. Cumulative frequency curve and its use in classification

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. Additional goals may emerge in particular application domains. Thus, in demographic applications it may be necessary to minimise differences between the classes in total population or total area. The analyst needs such tools that would allow her/him to balance between these goals in search of an acceptable compromise solution.

A visual representation of the statistical value distribution can help the analyst to meet the statistical criteria. Yamahira, Kasahara, and Tsurutani (1985) suggested that classification could be supported by a frequency histogram. However, a histogram represents *results* of prior classification and therefore can hardly serve as a *tool* for it. The dynamic classification tool of CommonGIS includes a dot plot, or point graph. This kind of graph does not require prior classification and is well suited for including in the interactive classification device due to its modest requirements to screen space. On the other hand, an obvious problem is overlapping of point symbols that obscures the understanding of the distribution. Slocum (1999) uses so called dispersion graphs (for illustration purposes) in his discussion of the existing classification methods. This representation is based on classification into a large number of classes. Values fitting in a class are shown by dots stacked at the class position. Since dispersion graphs already involve classification, they are not so good as tools for producing other classifications.

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. The method of construction of the ogive is demonstrated in Figure 1. The curve represents the distribution of values of the attribute “Number of cars per capita” over districts of Leicestershire (this and further examples refer to the Leicestershire sample of the 1991 census data available at the URL http://www.mimas.ac.uk/descartes/).

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.

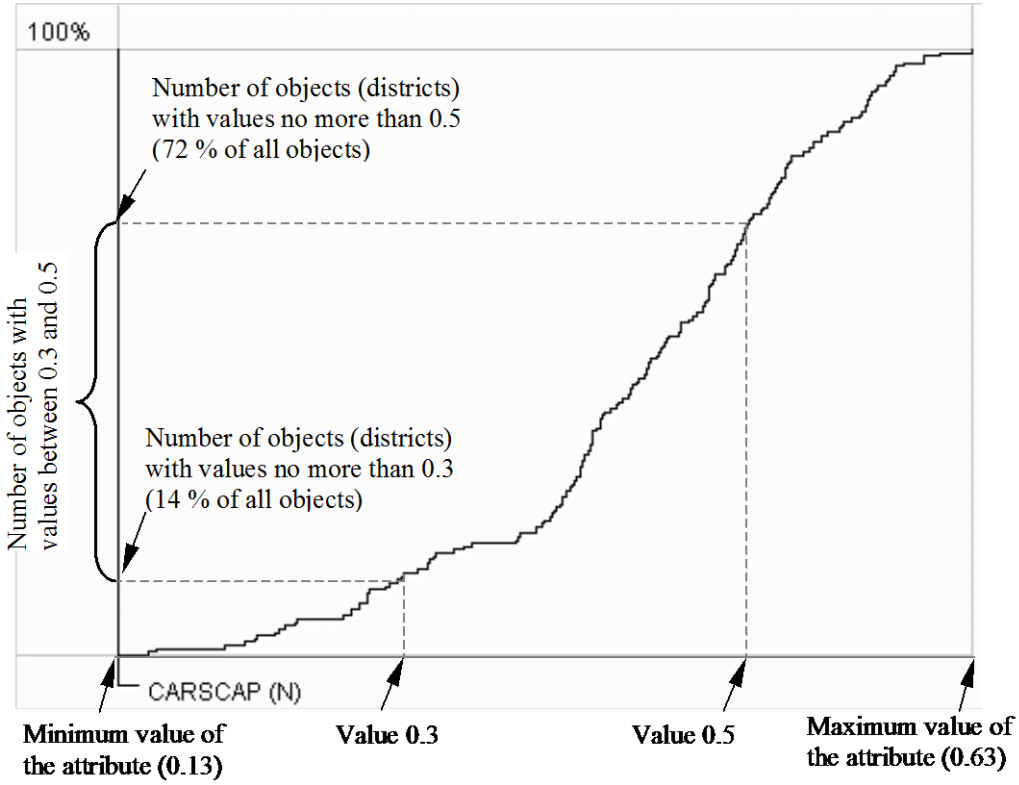

Figure 1. A cumulative frequency curve.

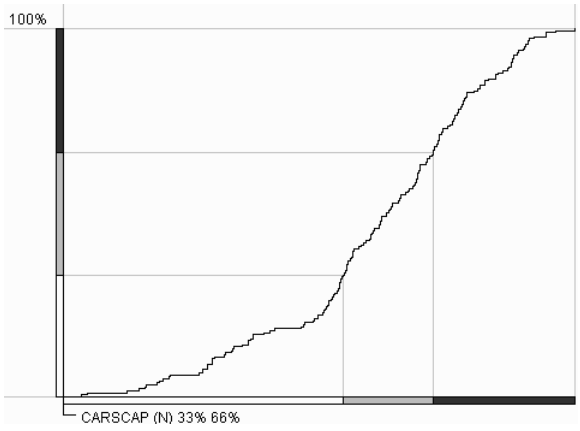

Figure 2. Representation of classes on a cumulative frequency curve display. Values of the attribute "Number of cars per capita" are divided by 2 breaks (0.404 and 0.491) into 3 classes with approximately equal sizes (33%, 33%, and 34% of the whole set). The positions of the breaks on the vertical axis are indicated below the graph: 33% and 66% (of the total number of the classified objects).

It is important that the cumulative frequency curve does not require prior classification. However, it *can* represent results of classification by means of additional graphical elements, and we used this opportunity in the latest extension of CommonGIS. Thus, the horizontal axis of the graph may be suited to show class breaks. In the interface adopted in CommonGIS (Figure 2) we

use for this purpose segmented bars with segments representing the classification intervals. The segments are painted in the colours of the classes. The positions of the breaks are projected onto the curve, and the corresponding points of the curve are, in their turn, projected onto the vertical axis. The division of the vertical axis is also shown with the use of coloured 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 class breaks shown in Figure 2 divide the whole set of objects into 3 groups of approximately equal size that is demonstrated by the equal lengths of the bar segments on the vertical axis.

The overall interface for classification is shown in Figure 3. Besides the cumulative curve display exposing statistical characteristics of the current classification, it includes a map showing geographical distribution of the classes. For the use of the cumulative curve as a tool for classification it is important that its display immediately reacts to any changes of the classes, as well as the map does. The user can change class breaks by moving the sliders (double-ended vertical arrows) along the slider bar (on the upper right of the window). In the process of moving the slider the map and the cumulative curve graph are dynamically redrawn. In particular, changing are the relative lengths of the coloured bars on the axes and the positions of the projection lines. Clicking on the slider bar introduces a new class break, bringing a slider close to another slider results in the corresponding break being removed. The map and the cumulative curve display immediately reflect all these changes.

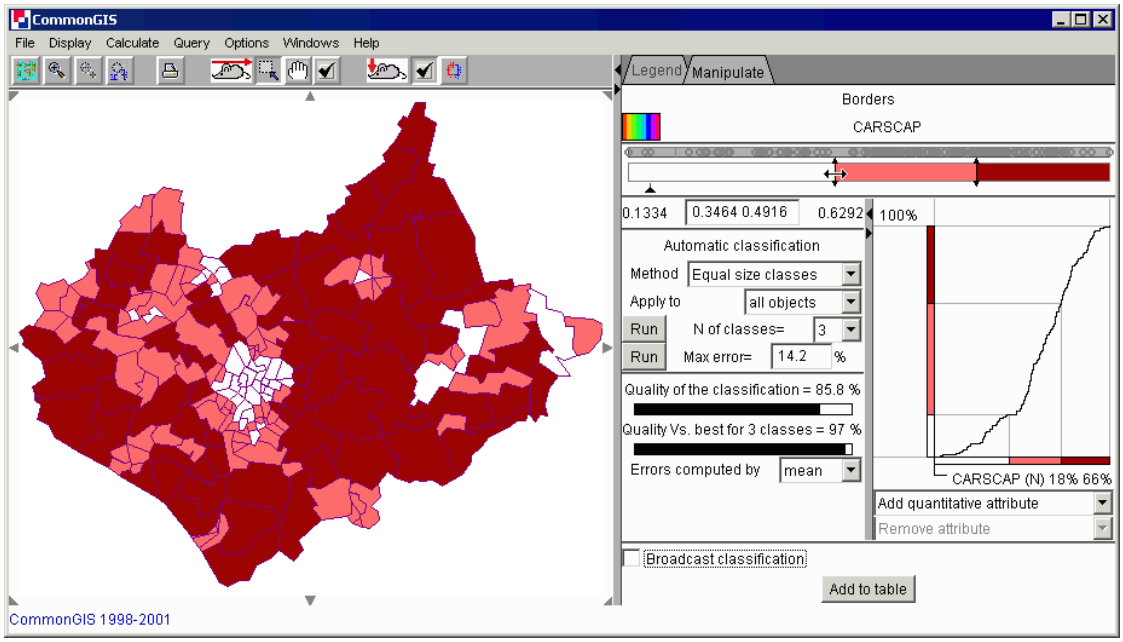

Figure 3. The interface for classification provided in CommonGIS allows the user to account both for statistical and for spatial distribution of values.

The use of the map and the cumulative frequency graph within the tool for classification allows the user to balance between the statistical and geographical criteria. For example, after looking at the graph shown in Figure 2 the user may wish to move the lower break from the area of clustering (indicated by a steep segment of the curve) to the horizontal segment on the left of it. This operation will improve the classification from the statistical perspective. At the same time the user can see how this affects the pattern visible on the map. When there is a break on a gentle slope, it may be reasonable to try to swing it around this position in search of a simpler spatial

pattern. It is also possible to focus during break movement on the vertical axis of the graph and produce classes with desired relative sizes (numbers of objects).

3 Generalised cumulative curves

In particular application domains additional classification criteria may come into play in combination with the statistical and geographical ones. For example, in demographic analyses it may be important to produce classes of districts that do not differ too much in total number of population that lives in them.

It is possible to generalise the idea of the cumulative frequency curve and to build similar graphs summarising values of arbitrary quantitative attributes. Examples of such attributes are area, population number, gross domestic product, number of households, etc. A generalised cumulative curve is built in the following way. Let the horizontal axis correspond to attribute A and the vertical to attribute B. The curve matches each value *x* of A with the sum of values of B computed for objects with the values of A being less than or equal to *x*. So, the maximum value of A will correspond to the total sum of values of B for all the objects of the sample. Let, for example, the districts of Leicestershire be classified according to the number of cars per capita, and a generalised cumulative curve represent the attribute "Total population". Then the vertical position corresponding to *x* cars per capita would reflect the total number of population living in districts with no more than *x* cars per capita.

The classification tool of CommonGIS allows the user to add a generalised curve for any quantitative attribute to the cumulative frequency curve display. The curves are overlaid, i.e. drawn in the same panel (see Figure 4). To be better discriminated, the curves differ in colour. The horizontal axis is common for all of them. The vertical axes are shown beside each other on the left of the graph. Each of the vertical axes is divided into the same number of segments, but positions of the breaks are, in general, different. This is clearly demonstrated in Figure 4. It shows that 34% of all districts fit in the lower class of the classification. They occupy only 10% of the total area but contain 50% of total population.

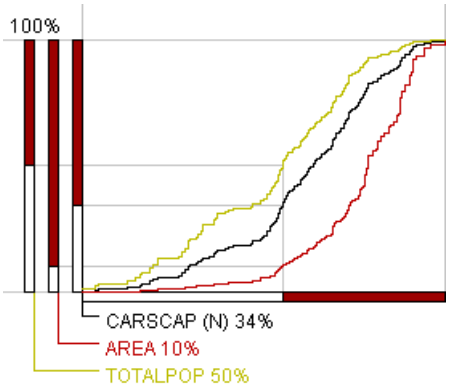

Figure 4. Generalised cumulative curves are built for the attributes "Area" and "Total population". The classification is done on the basis of the attribute "Number of cars per capita".

Having such a tool, it is easy to account in classification for such criteria as even distribution of population among the classes, or approximately equal total areas occupied by the classes, or other specific criteria that may emerge in this or that application domain. Thus, to make classes

approximately equal in total population, the analyst should focus in the process of slider movement on the axis corresponding to total population and try to position the sliders so that the axis is divided into segments of equal length.

Besides this opportunity, generalised cumulative curves may be used for exploring relationships between various characteristics of the classified objects. Let us demonstrate this on the example of exploration of unemployment in Leicestershire. We used the attributes "Number of unemployed" and "Total population" to calculate percentage of unemployed in total population in each district. Then we took this new attribute as the basis for classification. The classification tool showed us that proportion of unemployed in population varies from 0.9% to 13.62%. We considered values above 4% to be very high and wondered in how many districts this threshold is exceeded and where these districts are located. We entered 4 as a class break and in this way divided all districts into two classes: with up to 4% of unemployed persons in population and with more than 4%. The cumulative curve display showed us that only 18% of all districts fit in the upper class (Figure 5). The map shows a vivid spatial cluster of such districts in the centre of the area. It is seen that these districts occupy a rather small part of the whole area. However, when we selected the attribute "Total population" for representation on the cumulative curve display, we found that the districts with high unemployment contain 33% of the total population of Leicestershire.

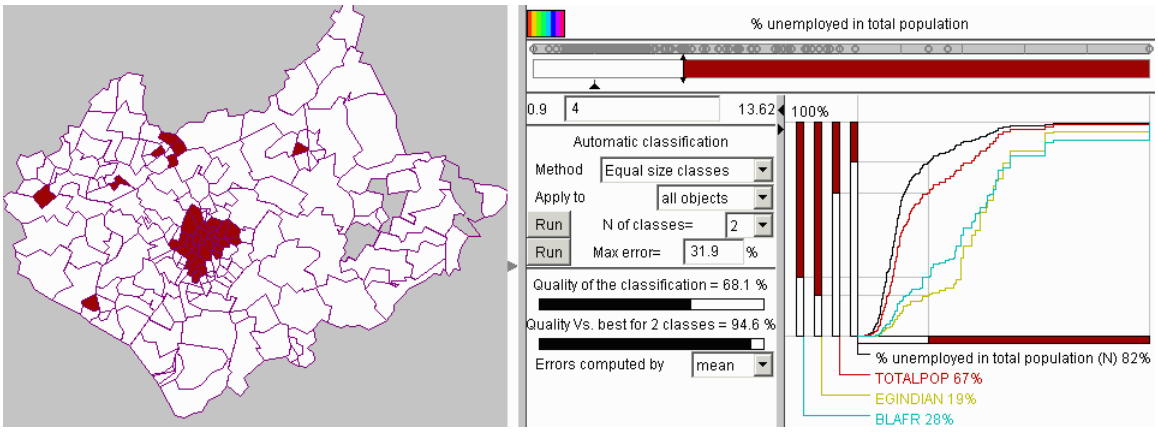

Figure 5. The use of generalised cumulative curves for exploration of unemployment in Leicestershire.

We became interested whether there is a link between unemployment and distribution of national minorities. We added to the cumulative curve display the attribute BLAFR representing total numbers of people originating from Africa by districts. In Figure 5 it is vividly seen that the curve for this population group radically differs from that for the whole population. It is also seen that the axis corresponding to this attribute is divided in quite a different proportion than those for the frequency and for the whole population. Only 28% of people with African origin live in districts with lower unemployment and, hence, 72% live in districts with more than 4% unemployed in total population. The difference is even more dramatic for people originating from India (represented by the attribute EGINDIAN). One can see that 81% of these people live in the areas with high unemployment.

We continued our investigation of unemployment by moving the class break so that the population was divided into two equal parts. Figure 6 shows the result of this operation. The new value of the class break is 2.89. This means that 50% of total population lives in districts with

more than 2.89% of unemployed. These districts, as it is seen from the map, constitute a rather small part of the whole territory of the county. Hence, the population density in them is higher than in the rest of the districts. Apparently, these are mainly urban districts. The map shows also that the districts with higher unemployment are spatially clustered. The national minorities considered above are now distributed between the classes of districts in the following way: only 14% of Indians and 18% of Africans live in the districts with lower unemployment, and, hence, 86% and 82%, respectively, live in the areas with more than 2.89% unemployed in population.

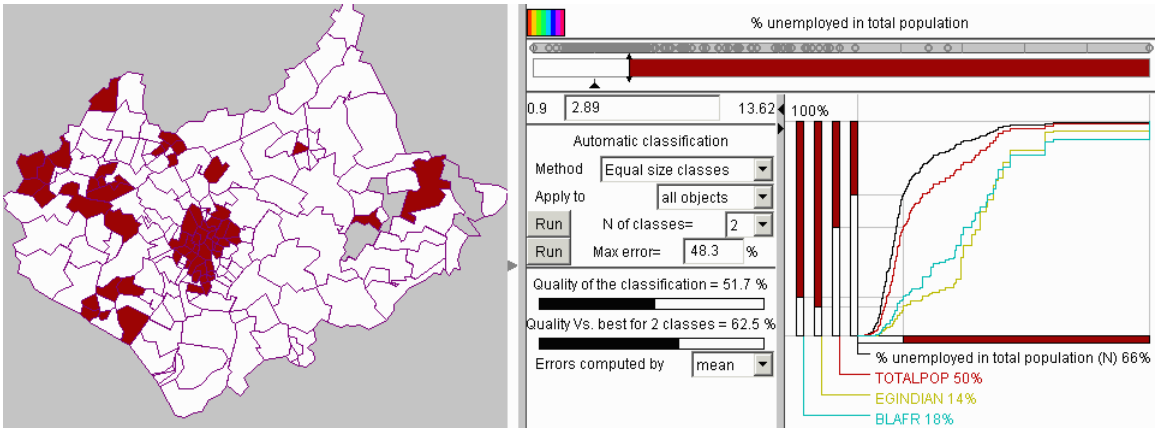

Figure 6. The cumulative curve display was used to divide the districts into two classes with equal total population.

This example analysis demonstrates that generalised cumulative curves can not only facilitate classification with multiple criteria involved but also to reveal significant relationships in data. However, it should be borne in mind that this technique is suitable only for attributes the values of which can be summed up over the set of objects they refer to. For example, it would be wrong to apply it to percentages, averaged values, rates, values per capita etc.

The implementation of the interactive classification tool based on the use of cumulative curves is done in Java. This allows the use of it on different platforms and in the WWW. The new version of the system CommonGIS that includes the tool described can be run in the WWW at our homepage.

Conclusion

Interactive, dynamic classification can be a valuable instrument in exploratory analysis of spatially referenced data. Representation of classes on a map by colouring facilitates perception of patterns of spatial distribution. Therefore we included a dynamic classification tool into the array of exploratory facilities offered by our system CommonGIS.

In classification of geographical objects according to values of a numeric attribute it is necessary to take into account peculiarities of both spatial and statistical distributions of values. In search for a suitable representation of the statistical distribution we studied the properties of the cumulative frequency curve and found it to be a good solution. This representation allows a two-way use. On the one hand, one can visually evaluate statistical characteristics of a given classification. On the other hand, one can produce classifications with desired statistical characteristics. In our implementation a cumulative curve display is included in the interface for

classification together with direct manipulation controls and a map showing the results of classification in the geographical space. In this interface the user can gradually shift class breaks and immediately observe the effect on the map and on the cumulative curve display. This dynamic link between the components of the interface allows the user to evaluate "on the fly" lots of variants of classification from the perspective of satisfaction of geographical and statistical criteria and to arrive eventually at a good compromise solution.

In the process of exploration of the properties of the cumulative frequency curve we came upon an idea that this representation could be extended to arbitrary attributes that allow summing over a set of objects. Just as the frequency curve accumulates the number of objects, the generalised curve would accumulate the values of an attribute. The use of such curves in classification offers additional opportunities. One of them is accounting in classification for such criteria as even distribution of population or area among classes. Another opportunity is investigation of relationships between the attribute used for classification and various quantitative characteristics of the objects being classified.

We believe that inclusion of cumulative curves in the tool for classification available in CommonGIS significantly increases its exploratory value. Thus, this tool has got a high appraisal of an expert in statistics and statistical graphics and a professional in analysis of geographic information, also with a solid statistical background. However, we have a concern that the users with less expertise in statistics may find cumulative curves difficult to understand. We plan to perform experiments in order to evaluate how people can handle the classification tool and how much time they need to comprehend the cumulative curve display and learn to use it.

References

1. Andrienko, G. and Andrienko, N. (1998) Dynamic Categorization for Visual Study of Spatial Information. *Programming and Computer Software*, **24** (3), pp.108-115
2. Andrienko, G. and Andrienko, N. (1999) Interactive maps for visual data exploration. *International Journal Geographical Information Science*, **13** (4), pp.355-374.
3. Bertin, J. (1965/1983) *Semiology of graphics. Diagrams, networks, maps.* The University of Wisconsin Press, Madison WI.
4. Egbert, S.L. and Slocum, T.A. (1992) EXPLOREMAP: an exploration system for choropleth maps. *Annals of the Association of American Geographers*, **82**, 275-288
5. MacEachren, A.M. (1994) *Some Truth with Maps: A Primer on Symbolization and Design* (Washington: Association of American Geographers)
6. Robinson, A.H., Morrison, J.L., Muehrcke, P.C., Jon Kimerling, A., and Guptil, S.C. (1995*) Elements of cartography.* Wiley, New York
7. Slocum, T.A. (1999) *Thematic Cartography and Visualization.* Prentice-Hall, New Jersey
8. Yamahira, T., Kasahara, Y., and Tsurutani, T. (1985) How map designers can represent their ideas in thematic maps. *The Visual Computer*, **1**, 174-184