Exploratory Analysis of Spatial Data Using Interactive Maps and Data Mining

N. Andrienko, G. Andrienko, A. Savinov, H. Voss, and D. Wettschereck

ABSTRACT: We present new methods for analyzing geo-referenced statistical data. These methods combine visualization and direct manipulation techniques of exploratory data analysis and algorithms for data mining. The methods have been implemented by integrating two hitherto separate software tools: Descartes for interactive thematic mapping, and the data mining toolbox Kepler. In using these tools, data analysis may proceed as a steady interaction between visual inspiration and insights gained from mathematical–statistical calculations. After introducing the various components of the methods and tools, the paper guides the reader through in-depth examples of using the tools in the context of analysis of urban demographic data. In particular, it is shown how geography-based classifications of urban districts can be related to available thematic characteristics by applying the data mining algorithms *classification tree derivation, attribute weighting,* and *subgroup discovery.*

KEYWORDS: Exploratory data analysis, interactive maps, spatial data mining

Introduction

Ever since the pioneering work of Tukey (1977), the importance of exploratory data analysis (EDA) has been widely recognized in modern statistics. The goal of exploratory data analysis is to gain understanding and seek information from hitherto untouched or insufficiently understood data, i.e., to penetrate into relationships, patterns, and trends hidden inside data. This differentiates EDA from more traditional techniques of statistics that typically check a priori hypotheses and are therefore characterized by Tukey as "confirmatory." Exploratory data analysis and traditional statistical analysis methods are natural partners in any data investigation: in the course of exploratory data analysis, an analyst develops plausible hypotheses that can be confirmed or refuted using confirmatory techniques.

Techniques of EDA are mostly based on *data visualization*, i.e., the graphical presentation of data in ways that prompt the discovery of important traits and relationships. Paper and pencil that were initially used for these purposes later gave way to graphical computer screens and software. Computers enabled features of graphical presentations that are now considered indispensable for EDA: high user interactivity, allowance for various transformations, and multiple, dynamically linked views which make it possible for changes in one display to be automatically propagated to all other views.

An important category of data dealt with in statistics is the category of *spatially referenced data*. Many statisticians who developed techniques for EDA have been concerned about proper ways of visualizing such data. McDonald (1982) suggested the idea of parallel viewing of geographical data on a map and on a scatter plot. In order to link the two displays, regions in the map were painted in the same colors as corresponding points in the plot. Later this idea was further developed and implemented in various ways (see, for example, Buja *et al.* 1991; MacDougall 1992; Symanzik et al. 1996; Cook *et al.* 1997). The notion of EDA and data visualization has spread from the realm of statistics to cartography where the concept of "geographic visualization" emerged (DiBiase 1990, MacEachren 1994, MacEachren and Kraak 1997). Lately, significant effort has been devoted to putting the power of cartographic methods of representation into service in spatial data exploration (Egbert and Slocum 1992; Dykes 1997; Andrienko and Andrienko 1999a).

A quite different approach to revealing significant relationships and peculiarities in data is taken in the research area called "knowledge discovery in databases" (KDD). Knowledge discovery in data-

N. Andrienko, G. Andrienko, A. Savinov, and **H. Voss** are part of the AiS.KD Knowledge Discovery Research Team [http://ais.gmd.de/KD/] at the German National research Center for Information Technology, Schloss Birlinghoven, Sankt-Augustin, D-53754 Germany. **D. Wettschereck** is at the University of Applied Sciences Bonn-Rhein-Sieg, Department of Applied Computer Science. E-mail: <gennady.andrienko@gmd.de>; http://borneo.gmd.de/and/. Tel: +49-2241-142486. Fax: +49-2241-142072.

bases is the non-trivial process of identifying valid, novel, potentially useful, and understandable patterns in data (Fayyad *et al.* 1996). This process consists of a number of steps such as data collection, data pre-processing, data analysis (called data mining in KDD), evaluation of results and, finally, the implementation of results. Typically, parts of the process or the entire process must be executed more than once with varied parameters in order to achieve best results. Numerous algorithms for the data mining step have been proposed (Fayyad et al. 1996).

In exploratory data analysis, it is the task of the human analyst to uncover important characteristics of the data. In KDD, the goal is to develop methods for the automatic extraction of knowledge from data. In fact, KDD may be regarded as a kind of exploratory data analysis, because data mining techniques are suggested for quite the same purposes as are the conventional EDA tools based on data visualization—i.e., for acquiring knowledge from previously unexplored data.

Researchers in the area of KDD have paid some attention to spatially referenced data, as specialized algorithms for processing such data have been developed (see, for example, Openshaw et al. 1987; Gebhardt 1997; Koperski et al. 1998). The task of these algorithms is to account, in some way, for the spatial aspect of the data: relative positions, adjacency, distances, and directions. To allow for processing by a data mining method, the spatial aspect has to be represented in a discrete, symbolic form appropriate for machine processing. Thus, Gebhardt (1997) assumes that spatial information is given in the form of a neighborhood matrix, while Koperski et al. (1998) use spatial predicates and functions. Usually, extensive computations are needed for obtaining such representations from digitized spatial data. For each particular kind of spatial relationship, a special algorithm has to be developed; for example, Zhang and Griffith (1997) describe an algorithm for deriving a neighborhood matrix from ArcInfo™ coverage files. A problem with this approach is that an exhaustive symbolic representation of *all* spatial information is never achieved. In general, this appears to be impossible because space is a continuous and multifaceted phenomenon. Typically, only a small part of spatial relationships existing in a data set is symbolically encoded and utilized in data analysis.

An appropriate visual representation of spatial data, such as a map, can be isomorphic to space and thus capable of preserving all spatial relationships. This representation is, however, only perceivable by human eyes and can therefore be used by human analysts only. Although the human eye can immediately grasp most spatial properties and relationships properly reflected in a map, our analytical capabilities are very limited in terms of the volume of data that can be analyzed and the complexity of the (relevant) information that may be hidden in the data.

Because neither data visualization nor knowledge discovery in databases alone can provide a universal approach to exploration of spatial data, one would expect a benefit from combining the two methods. Such a combination may compensate for deficiencies of each method and, possibly, bring about a synergy of the methods. In combinations with geographic visualization, one can try to employ not only data mining algorithms specifically designed for spatial data but also general KDD techniques. To make these techniques applicable, spatial information has to be encoded in a suitable symbolic form, e.g., as values of attributes. The encoding may be accomplished by utilizing interactive software for geographic visualization. Another important role of interactive maps is to provide a spatial context for the results of data mining that is essential for these results to be properly interpreted.

The remainder of our paper is organized as follows: In the next section, an integrated data mining/visualization framework is introduced, while the following section describes selected, widely used, data mining methods. In the last two sections, we use examples to demonstrate how exploratory data analysis may be performed with the use of the integrated system. In all the examples we use demographic data for administrative districts of the city Bonn, which were kindly provided by the Bonn city administration.

Framework for Integrating Data Mining and Interactive Visual Exploration

We have built an integrated environment for exploratory analysis of spatial data using two existing systems: Descartes (Andrienko and Andrienko 1997; 1999a) for geographic visualization and Kepler (Wrobel et al. 1996) for data mining. Descartes automatically selects appropriate map symbols based on characteristics of user-chosen data and supports various interactive manipulations of map displays that can help reveal important features of a spatial distribution of data. Descartes also enables some data transformations that are productive for visual analysis, such as map-

supported discretization of numeric attributes. It offers a convenient graphical interface for outlier removal and an easy-to-use tool for generation of derived variables by means of logical queries and arithmetic operations on existing variables.

Descartes has been successfully applied to various statistical data sets. For example, it has been used to visualize results of elections in the city of Bonn, i.e., proportions of votes for different parties (http://borneo.gmd.de/and/java/iris/app/elect/index.html). In the United Kingdom, Descartes has been used within the KINDS (Knowledge Interfaces to National Data Sets) project (Andrienko et al. 1999) for presentation of British Census 1991 data. This system, along with the census data, is accessible through the Internet (http://www.mimas.ac.uk/descartes/) for the academic community of the UK. Within the EU-funded CommonGIS project (http://commongis.jrc.it/), Portuguese census data are being analyzed using Descartes. Currently, Eurostat, JRC (Joint Research Center of the European Community) and GMD (German National Research Center on Information Technology) are developing a Descartes-based demonstrator for the visualization of data from the NewCronos database (Guittet et al. 2001). Various demonstrators of Descartes are available on the Internet at the URL http://borneo.gmd.de/and/java/iris/.

Kepler is a KDD support engine that includes a variety of data mining methods. Kepler also contains tools for data pre-processing, access to databases, and querying, and it is capable of graphical presentation of several types of data mining results such as classification trees and rules or subgroups. It should be noted, however, that the methods currently available in Kepler are not specifically designed for spatial data. By means of combining with Descartes we sought to bring the power of these techniques to analysis of spatial information. The integrated environment we developed supports the following data exploration scenario.[1]

First, an analyst uses the mapping facilities of Descartes to preview the data to be studied, that is, to observe characteristics of the territory and the distribution of spatial objects and values of attributes attached to them. Then the analyst uses the interactive map manipulation tools of Descartes to encode spatial information in a form suitable for processing by the data mining procedures. For example, the analyst may divide spatial objects according to their locations into classes such as "North," "Center," and "South." In the next step, the relevant data (including both original and interactively produced attributes) are submitted to Kepler for further analysis. Within Kepler, the analyst starts a data mining analysis and then visualizes and interprets the results of this analysis. In many cases, viewing and interpretation of data mining results obtained in Kepler is supported by interactive maps within Descartes. The analyst may also return to the interactive maps in order to change previously made selections and encoding and to re-run data mining procedures. This iterative process is consistent with the very nature of exploratory data analysis, where the analyst has no initial knowledge about the data and, hence, cannot select from the beginning the "right" method and the "right" settings that will necessarily expose significant features of the data.

Visual presentation of results of data mining may significantly help the user in the interpretation of these results. Although Kepler can display graphically some types of output from data mining procedures, it is incapable of representing the spatial aspect of data. Thus, many data mining methods generate logical expressions describing groups of objects. In the case of spatially referenced data these groups consist of spatial objects, for example, "houses with 3 bedrooms and more than 12 acres land," a method might find that such houses cannot be sold in less than two weeks. It is essential to give the user an opportunity to locate objects of any such group on a map. Since Kepler has no mapping facilities of its own, it is the link with Descartes that makes this possible.

Two facts that can be observed in most real world applications complicate the presentation of KDD results on a map. First, due to the complex nature of most applications, data mining methods typically find a large number of potentially interesting groups of objects. Second, depending on the mining method employed, they may overlap, i.e., two groups with syntactically different descriptions may describe the same objects ("districts with high proportions of young people" vs. "districts near the university"). Therefore, in most cases it is difficult to represent the whole output of a data mining method in a single map, while keeping the map legible and useful for analysis. On the other hand, it may be inappropriate to show each group on a separate map, as the user may not be able to make any reasonable use of the resulting abundance of maps. A more feasible solution is to allow the user to interactively select a desired group and to display the spatial distribution of this group. For this purpose, the system can either automatically gen-

[1] The architecture of the link between Descartes and Kepler is described in Andrienko and Andrienko (1999b).

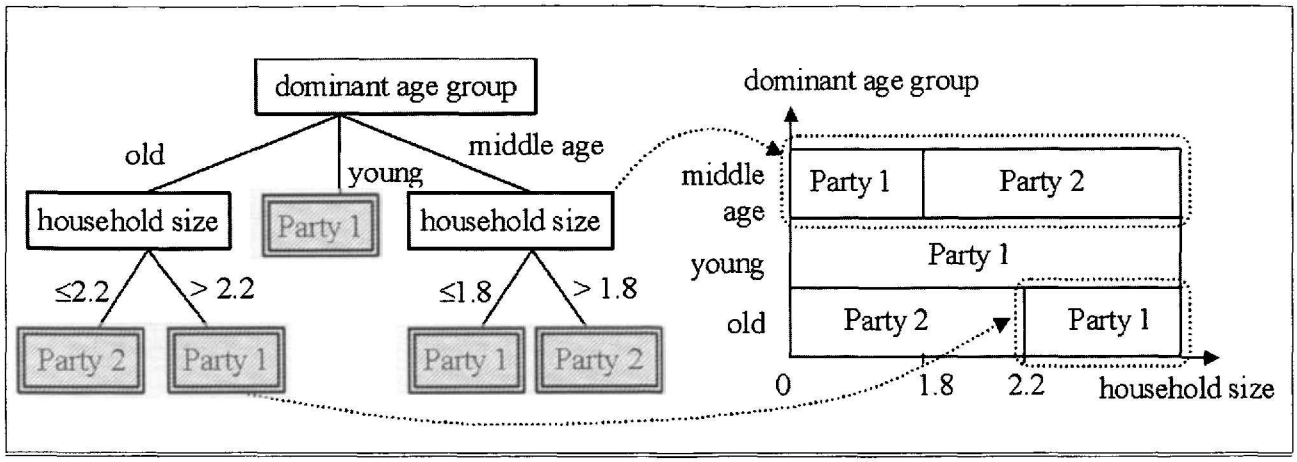

Figure 1. An example of a decision tree.

erate a new map showing this group or dynamically change an existing map display so that the objects constituting the group are drawn differently from the rest of the objects (highlighted).

In our implementation we have chosen the second option. When the user has some data mining results displayed graphically in Kepler and points with the mouse on a graphical element representing a group of geographical objects, the objects belonging to that group are highlighted in all the map displays currently open in Descartes. The dynamic link between the systems is bi-directional: when the user clicks on some spatial object in a map, the graphical element(s) representing the group(s) it fits in are highlighted in the presentation of the data mining results in Kepler. This link is essential for an appropriate interpretation of data mining results as it provides the missing spatial reference.

Data Mining Methods Used in the Experiments

In our experiments we used three of the most widely used data mining methods:

- A method that derives *classification trees* (often also called *decision trees*);
- A method that estimates the relevance of given attributes in predicting the values of an independent attribute (*attribute weighting*); and
- A method discovering interesting *subgroups* of objects.

These methods were chosen from the large variety of existing data mining methods for their relative simplicity and their direct applicability to the tasks described in this paper.

Classification Tree Derivation

In data mining terms, classification is prediction of the value of one attribute of a given object from a set of known attributes of the same object. For example, from a given set of symptoms and some patient data, a doctor can judge (predict) whether a patient suffers from a given disease. Classification could also be employed to predict whether people in a given area tend to vote for one of two parties (see the example in Figure 1). A classification tree is a specific representation of a set of rules that derive the value of one attribute (the *class* to be predicted) from the values of a set of given attributes. The term *tree* stems from the fact that a classification tree resembles a real tree when drawn, with the exception that it has only a single root. The root usually has at least two children called nodes. Each node may in turn have several children. Nodes that have no children are called leaves. Each node of a classification tree tests the value of some attribute (the root node in Figure 1 tests the *dominant age group* of a district). An object to be classified is passed through the tree; depending on the outcome of the test at each node it passes, the object is passed on to the child that satisfies the test until a leaf is reached (districts with "dominant age group = old & household size <= 2.2" are passed to the leftmost leaf). Each node therefore describes a conjunction of the conditions of its parent nodes. This conjunction characterizes a certain group of objects.

The C4.5 classification derivation algorithm (Quinlan 1993) is one of the best known, efficient, and widely used KDD algorithms. The goal of the algorithm is to find a near-optimal order of tests (tree nodes) in order to construct a classification tree with as *pure* leaves as possible, while minimizing the total number of nodes. A pure leaf is a leaf

that contains as many objects of a given class (for example, voters for Party 1) as possible and as few objects of other classes (voters for all other parties) as possible. The decision tree is built recursively, starting from the initial (root) node corresponding to the whole set of objects and then executing the same procedure for each node down to the leaves. At each step the attribute is chosen for the test that produces the purest child nodes. For numeric attributes, all different values of the attribute are evaluated and the best split value is chosen (2.2 for the left subtree and 1.8 for the right subtree). The procedure continues until either only objects from one class are contained in the leaf or no test can be found that would produce a purer division of the objects. In real world applications, leaves therefore often contain exceptions.

Once a decision tree has been generated, it can be used for two main purposes: description and prediction. In the first case, the generated structure is presented to human analysts or decision makers to assist in the explanation and better understanding of the problem domain. In the second case, the structure is used to predict the unknown class of an object on the basis of known values of other attributes.

Attribute Weighting

Attribute weighting, also known as feature weighting (Wettschereck et al. 1997), is a family of algorithms that assign numeric weights to attributes according to their discriminatory effectiveness with regard to the given class attribute. Attributes that are the most informative, i.e., allow for better differentiation of classes, receive higher weights than attributes contributing less to class differentiation. One particular feature-weighting method included in Kepler and employed here is *mutual information*. The mutual information between two variables is the degree of reduction in uncertainty concerning the possible values of one variable that is obtained when the value of the other variable is fixed. If a feature provides no information (uncertainty reduction) about the class, the mutual information will be zero. If a feature completely determines the class, the mutual information will be proportional to the logarithm[2] of the number of classes (assuming examples from different classes have equal frequencies).

For example, if we have no further knowledge about any other attribute of a human being, then we have a 50 percent chance of correctly guessing his/her gender. However, if we know that the person plays American football, then we will guess with 95 percent confidence that he is male, assuming that we know that 95 percent of all football teams are male teams.[3] Hence, knowledge of "plays American football" will reduce our uncertainty about "Gender". The mutual information between "plays American football" and "Gender" is therefore greater than 0 and less than 1, depending on the actual probability distributions of "plays American football" and "Gender."

Subgroup Discovery

The key point in subgroup discovery methods is to discover subsets (subgroups) of objects in which the statistical distribution of the values of a selected attribute, called "target attribute," deviates maximally from the distribution of values of this attribute within the entire data set. For example, 80 percent of objects of a subgroup have value A of an attribute and 20 percent have value B, while in the whole data set the proportion is 40 percent to 60 percent. Because of this distribution difference, the subgroup is considered to be interesting. The number of possible subgroups is potentially extremely large (thus, from a set consisting of N elements 2^N different subsets can be selected) while only a few of them are really interesting. It is a general goal of subgroup discovery methods to efficiently find such interesting data subsets in the original data. Subgroups are described as a conjunction of conditions on non-target attributes (for example, gender = male & age > 40).

A significant part of a subgroup discovery method are evaluation criteria which determine what characteristics have to be taken into account when deciding whether a subgroup is interesting. The most frequently used criteria are deviation patterns (Klösgen 1996) which compare to what extent the distribution of values of the target attribute(s) over a subgroup is different from that in the whole data set.

Like the classification tree derivation methods, the methods for subgroup discovery can be applied to classes of objects (the classes are treated as values of the target attribute). The most substantial difference between these two groups of methods is that in subgroup discovery, a subset is considered to be interesting if the distribution of its objects among the classes differs maximally from the root distribution, while in classification tree derivation, it would only be interesting if one class signifi-

[2] This is derived from information theory as the number of bits required to encode the values of the class attributes.

[3] The 95 percent value was made up for the sake of the example.

cantly dominates the others. Subgroup discovery is therefore superior in applications where one is interested in learning about rare events. For example, highly interesting for a company might be a group of customers who are three times as likely to return as the average customers. Assuming that the average return rate is 10 percent, then C4.5 would never discover such a group, but a subgroup discovery algorithm will do so. Another case when subgroup discovery could be more appropriate is when an analyst is interested in finding subsets with unusually low proportions of objects of a particular class, e.g., districts that tend to vote either for Party A or Party B but not for Party C.

Example 1: Acquisition of Thematic Characteristics

As may be seen from the above description of a few representative data mining methods, many of them are applied to some classes of objects. The type of objects handled by these methods is irrelevant. The objects may be people, production processes, or spatial objects such as countries, districts, or earthquake epicenters. In the integrated system, whenever a Descartes user has a map on the screen with any classification of geographical objects, he or she can submit this classification to Kepler for running some of the available data mining methods. When the user submits this task to Kepler, the assignment of geographic objects to classes is retrieved from the map and sent to Kepler together with other available data about the objects (values of attributes). More details about the access from Descartes to the C4.5 method can be found in Andrienko and Andrienko (1999c).

Descartes offers various methods for classifying spatial objects. In the context of exploratory analysis of spatially referenced data, we are especially interested in the possibility of classifying spatial objects according to their locations and other spatial properties. For example, objects can be classified according to their relative positions with respect to other objects from the same set (north-south, center-periphery etc.) or to spatial relationships with other geographical features (closeness to sea, mountains, and roads). Such geography-based classification is a possible way to encode spatial information in a form that can be processed by general KDD algorithms.

Map-mediated Object Classification in Descartes

The classification of spatial objects according to geographical properties and relationships can be done in Descartes through direct manipulation of the map. Figure 2 shows the user interface of the map-based classification tool. The panel on the left of the map lists currently existing classes. The user may change the names and colors of the classes, or add and remove classes. For each class there is a radio button, the selection of which makes this class "active." Thus, in Figure 2 the class "South–West" is active. When the user clicks on an object in the map, the object is attached to the active class.

A classification made in this way can be transformed into a new attribute to be added to the existing table of thematic data about the spatial objects. When the user presses the button "Classes→table" in the lower left part of the window (see Figure 2), a dialog window appears which prompts the user to give a name to the new attribute and offers three options concerning the attribute's type: character (qualitative), numeric, or logical. By default, Descartes generates a qualitative attribute with class names taken as values. The user can replace the class names by other labels. To produce a numeric attribute, the user should provide a numeric value label to stand for each class. To derive a logical attribute, the user can assign to each class either the label T or F.

In our example, we classified the districts of Bonn into five geographical parts: Center, North-West, North-East, South-West, and South-East (Figure 2). Our objective was to check whether these parts differ in thematic properties, in particular, population structure.

Application of the C4.5 Method

To investigate the relationships of the defined classes with available thematic attributes characterizing the districts, we submitted the classes together with available thematic data to Kepler and ran the C4.5 method. The outcome of the method was a decision tree, a fragment (the top 3 levels) of which is shown at the bottom of Figure 3. Each node of the tree specifies a test of some attribute, and each branch descending from that node corresponds to one of the possible value subsets/intervals for this attribute. The test divides the objects coming to the node into subsets according to the values of the attributes and passes these subsets to the appropriate descending nodes. The colored segmented bar at the top of each node shows the distribution of the corresponding (sub)set of objects among the classes. The segments have the colors of the classes, and their lengths are proportional to the numbers of objects in the respective classes. The number of objects in each class,

absolute and in relation to the size of the whole (sub)set, is also shown below the bar.

The top node of the tree in Figure 3 contains a test of the attribute "Percentage (Age groups=18-30)" (i.e., percentage of people 18 to 30 years old in the total population) that is applied to the set of all districts of Bonn. This set is divided into two subsets according to the values of the attribute: districts with values equal or below 19.03 and those with values greater than 19.03. Remember that the C4.5 method tries to find interval breaks that divide the objects into groups approximating the given classes (ideally, coinciding with the classes). Thus, the break 19.03 divides the districts of Bonn so that one of the two resulting groups consists mostly of the districts of the central part (15 of 18 group elements belong to the class "Center") and includes almost all such districts (15 of 16). The information about this group can be seen in the right node on the second level of the tree.

As was already mentioned, graphical displays of data mining results in Kepler are linked with all maps and supplementary graphics in Descartes. The spatial distribution of objects at any node of the tree can be displayed on a map in Descartes. Thus, in order to view on the map the positions of the districts with high percentage of people between 18 and 30 years old, we clicked on the tree node representing this group of districts. This node is shown by the black outline in Figure 3. In response, Descartes marked the districts on the map by painting their borders in white color and putting small square boxes inside them (see the map in Figure 3). It is clear from the map that, indeed, the districts of the class Center (with only one exception) are characterized by the selected node of the tree, i.e., have more than 19.03 percent of young people. One can also see which three districts from other parts of the city fit in the same tree node.

By inspection of the node and its descendants we learn that two of these three districts, both belonging to the class North-West, can be separated from the rest of the group according to the values of the attribute "Percentage (Age groups=0-18)" (percentage of people aged between 0 and 18 years in the total district population). These two districts have more than 18.01 percent of children in their population, while in the other districts in the group the percentage of children is less than or equal to 18.01 percent. Again, the break 18.01 was found by the C4.5 algorithm in an attempt to fit the attribute-based division of objects to the given classification.

Another interesting node is the second from the left on the third level of the tree. This node represents the group of districts having no more than 19.03 percent of people aged from 18 to 30 years and more than 12.1 percent of foreigners in their population. This group contains 13 of 16 districts of the class South-East and only two districts from other classes.

In summary, the C4.5 method has exposed to us that the central and south-eastern parts of the city significantly differ in population structures from the other parts. The central part is characterized by an unusually high percentage of young people (i.e., between 18 and 30 years old), and the south-eastern part is distinguished by a very high percentage of foreigners.

Application of the Attribute-weighting Method

The attribute weighting method of data mining allows the user to find out on which characteristics (in terms of values of attributes) object classes differ most significantly. The method determines the relative effectiveness of attributes in differentiating classes and expresses this relative effectiveness through numeric weights of the attributes. It is essential that the attribute weighting method "weights" each attribute independently of others. This is different from the C4.5 algorithm where selection of attributes and breaks on lower levels of a tree depends on selections made on upper levels.

In our experiment, we applied the attribute weighting method to the geography-based classification of the districts of Bonn introduced above. The results, the computed weights of all attributes present in the database, can be seen in Figure 4. The attributes are arranged in the order of decreasing weight (i.e., effectiveness in differentiating the classes), with the weights represented graphically by lengths of bars.

The attribute weighting algorithm has found that the parts of Bonn differ most profoundly in "Percentage (Age groups=18-30)", which is consistent with choosing this very attribute as root predicate of the previously generated decision tree. Somewhat lower, but still rather high weights are assigned to the attributes "Percentage (Age groups=45-60)" (percentage of people 45 to 60 years old in the population) and "% Change 31.07.1996/31.12.1995" (percent change of population from December 31, 1995, to July 31, 1996).

We can verify these findings using Descartes, which allows us to see how the parts of Bonn are differentiated by the attributes that obtained the highest weights. The system can calculate and visually present summarized characteristics

of arbitrary object classes: minimum, maximum, mean, median, and quartile values of any of the attributes for each class. For example, Figure 5 presents summarized characteristics of the parts of Bonn in terms of the attributes "Percentage (Age groups=18-30)" and "Percentage (Age groups=45-60)". The bar charts inside the colored rectangles show the mean attribute values for the parts. The first (red) bar in each chart corresponds to the 18- to 30-year age group, and the second (blue) to the other age group. The exact values of the calculated summary characteristics can be retrieved by clicking in the rectangles. The "box-and-whiskers" plots below the rectangles show the statistical distribution of values of the attributes in each class. An explanation of the "box-and-whiskers" plot (sometimes also referred as "box plots") and its use in exploratory data analysis can be found in Tukey (1977).

In Figure 5 one can observe that the center of Bonn differs rather significantly with respect to the relative sizes of the two age groups from the other city parts, whereas the other parts are quite similar to each other. Thus, the attributes that have received the highest weights actually distinguish only the center from the periphery.

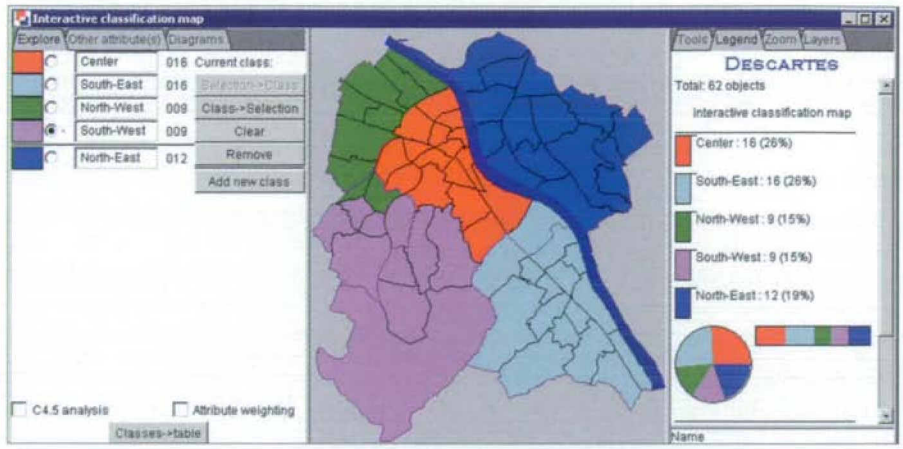

Figure 2. The map-based interface for interactive classification of spatial objects.

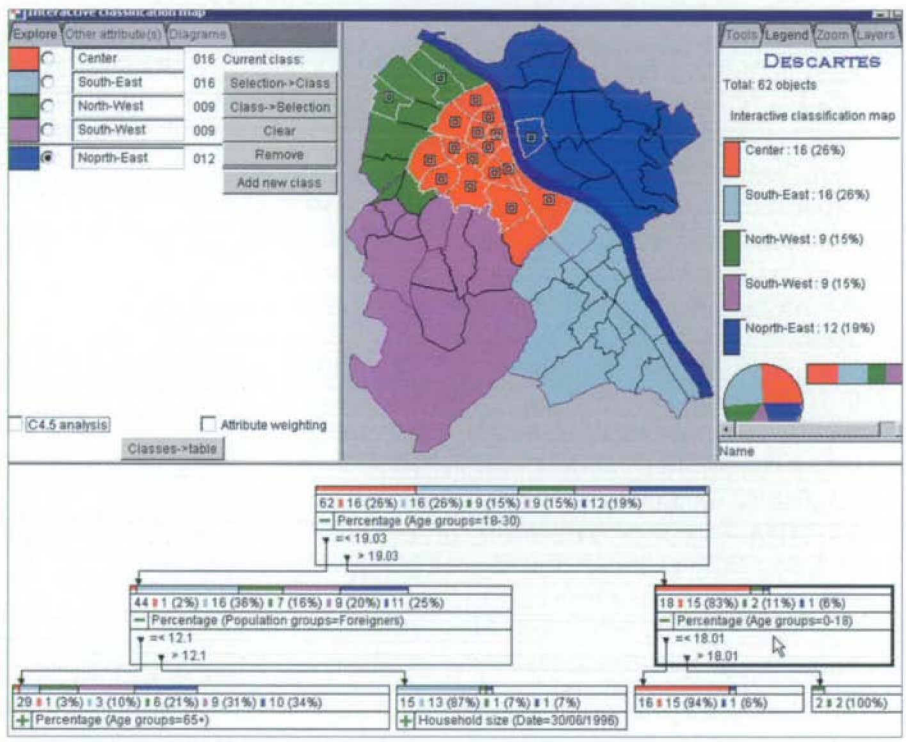

Figure 3. The decision tree derived for the classification of Bonn districts into five geographical zones. Marked on the map above the tree are the districts fitting in the currently selected tree node (enclosed in a thicker frame).

Let us conduct a similar investigation for the attribute "% Change 31.07.1996/31.12.1995". The summarized characteristics of the parts of the city in terms of this attribute are shown by bars and box plots at the top of Figure 6. The bars on the map below represent the values of the attribute associated with the individual districts. The upward orientation of a bar and the green color indicate increase of population, while the light blue bars directed downwards signify population decrease. The length of each bar is proportional to the amount of the change in population (in percent).

Both the map and the summary view above it expose a strong population decrease in the center of the city and an increase in the north–east. The mean attribute values for the remaining three parts of the city indicate a slight population decrease. However, the map shows considerable variation of values within these parts: in each of them there are both upward- and downward-oriented bars of substantially different lengths.

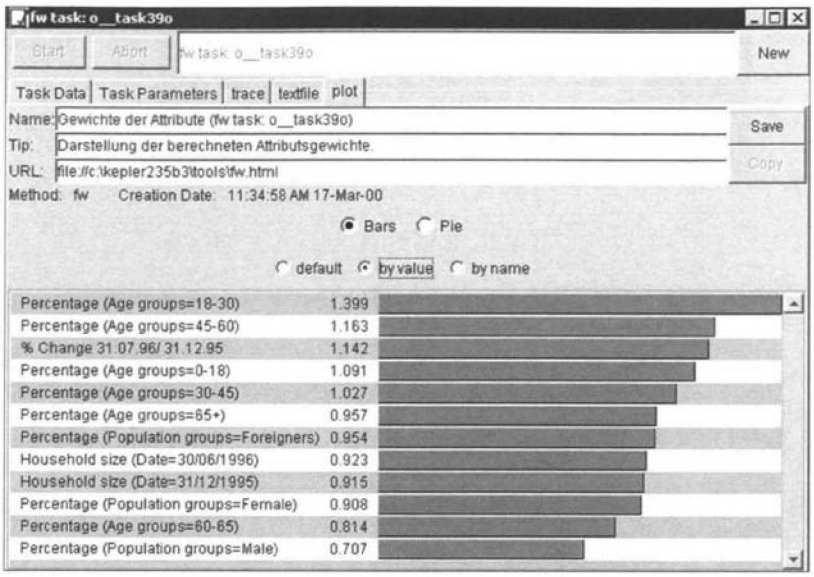

Figure 4. The results of an application of the feature-weighting algorithm to the geography-based classification of the districts of Bonn.

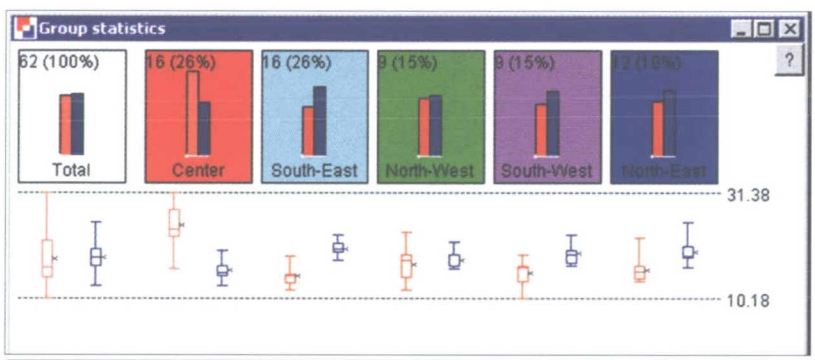

Figure 5. Characterization of city parts by proportion of people aged 18 to 30 years (red bars and box plots) and 45 to 60 years (blue bars and box plots).

To sum up, we have uncovered the following facts with the help of the attribute-weighting algorithm:

- Large differences between the city center and its periphery in the age structure of the population;
- Movement of population away from the center; and
- Population growth on the north-east side of the city.

Application of the Subgroup Discovery Method

Now we will consider what facts about the geographical parts of Bonn can be revealed using the subgroup discovery method. We assign the geographical classification to be the target attribute of the method. The task of the method is to find subgroups of objects such that the relative frequencies of values of the target attribute within a group differ significantly from those in the whole set of objects. Subgroups are defined in terms of values of the attributes present in our database: for example, "proportion of females is high and proportion of people aged 30 to 45 years is low."

The subgroup discovery method is able to generate subgroup definitions only on the basis of attributes with discrete value sets. This is an explainable limitation: an attribute with a continuous value range (e.g., numeric) allows for an infinite number of logical expressions that can be used for subgroup definition. Therefore, before applying the method, the user has to discretize the available numeric attributes, that is, split their value ranges into subintervals and make the system treat all values within a subinterval as the same value.

In our example, we discretized all the numeric attributes by dividing their value ranges into two subintervals—less than or equal to the median value and more than this value. With the exception of the attribute "household size," the original attribute values of the objects were substituted with the "low" and "high" labels. The median value of the latter attribute occurred very close to 2, and we decided to use the labels "LT 2" (less than or equal to 2) and "GT 2" (greater than 2).

The transformed data were submitted to a subgroup discovery method called MIDOS (Wrobel 1997), which found 10 "interesting" subgroup definitions. The frequencies of values of the target attribute (i.e. the geography-based classes) in the whole set of objects and in each subgroup are shown, both graphically and numerically, in Figure 7. The upper left pie chart corresponds to the whole set. From the numbers on the right of the pie it may be seen that 25.8 percent of the whole set, which consists of 62 districts, belong to the Center class, 19.3 percent to the North–East class, 14.5 percent to the North-West class, 25.8 percent to the South–East class, and 14.5 percent to the South–West class. The other diagrams and the figures on the right of them correspond to the

discovered subgroups. Definitions of the subgroups are given below the charts. The diagrams are drawn so as to allow for comparison of the subgroups with the whole set. For this purpose, a pie representing a subgroup was drawn inside of a pie representing the whole set. The size of the inner pie is proportional to the relative size of the subgroup. The angle sizes of the circle segments ("pie slices") show proportions of districts from corresponding classes (signified by colors) in the subgroup. These proportions can be compared to those in the whole set.

Thus, for example, the second diagram from the left in the upper row corresponds to the subgroup defined as "% 30-45 = low, HH size 95 = LT2" (i.e., proportion of people 30 to 45 years of age is low and mean household size in 1995 was less than or equal to 2). The figures on the right of the diagram show that this subgroup consists of 10 districts, and all 10 of them belong to the South–East class. The pie chart showing the frequencies of the classes has in this case a form of a circle painted in the color of the South–East class.

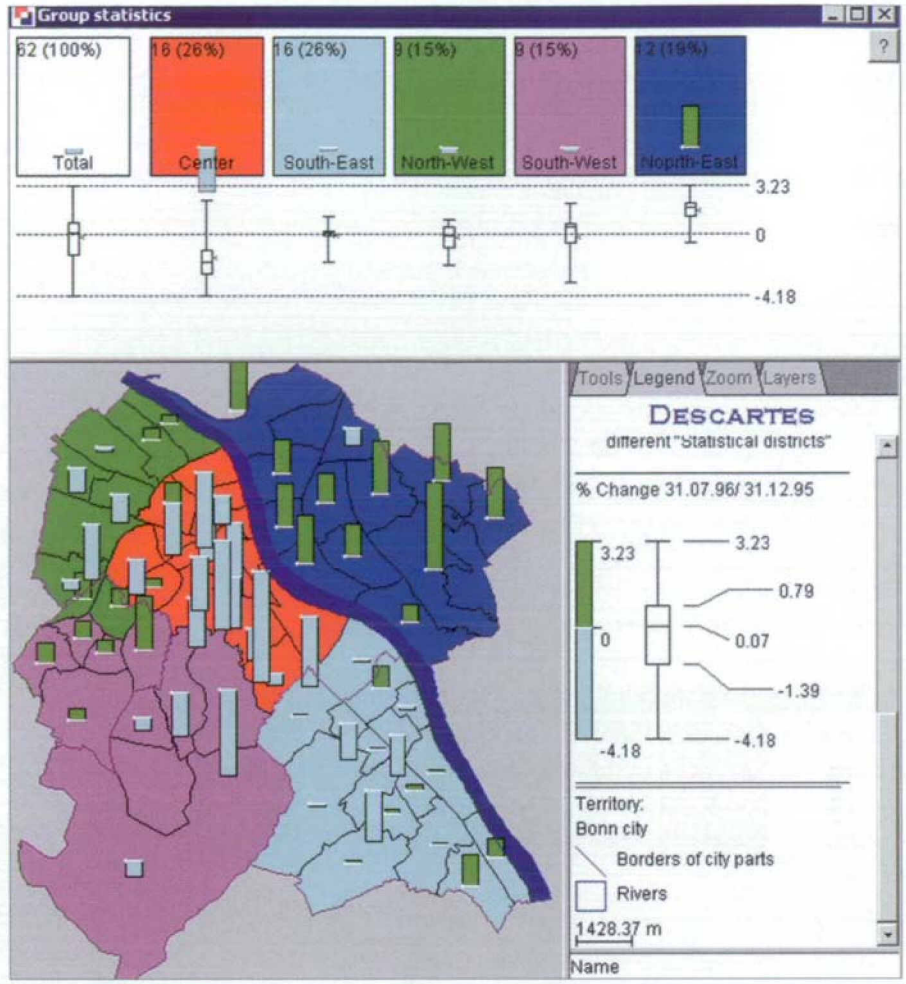

Figure 6. Relative changes of population from December 31, 1995, to July 31, 1996, for districts (shown on the map) and for the five geographical zones (summarized in the summary view above the map).

Cases when a sufficiently large subgroup contains only representatives of a single class are rather remarkable; these class members have common characteristics distinguishing them from objects belonging to other classes. Valuable facts about classes are provided by subgroups that consist mostly of members of one class but also include a few objects from other classes. For example, the subgroup third from the left in the second row consists of 17 objects, 13 of which (76.4 percent) belong to the Center class. The subgroup is defined as " %<18 = low, % 30-45 = high" (i.e., proportion of people less than 18 years of age is low and proportion of those 30 to 45 years old is high). We thus learn that 13 central districts have these peculiarities of population structure (note that the class "Center" contains a total of 16 districts).

All the detected subgroups can be divided into two categories: one distinguished by very high proportions of districts of the class South-East and the other with districts mostly from the class Center. This means that all the data mining methods have found that the districts comprising these two parts of Bonn have coherent and rather particular demographic characteristics distinguishing them from the rest of the districts. The other parts of the city, apparently, cannot be consistently characterized in terms of the available thematic attributes due to a large variation of characteristics within the parts.

In Figure 7 one can observe the following logical expressions in the definitions of the subgroups with remarkably high proportion of districts from the south–east:

- Low proportion of the 30-45 age group (subgroups 1, 2, 7, and 10, counting from the left

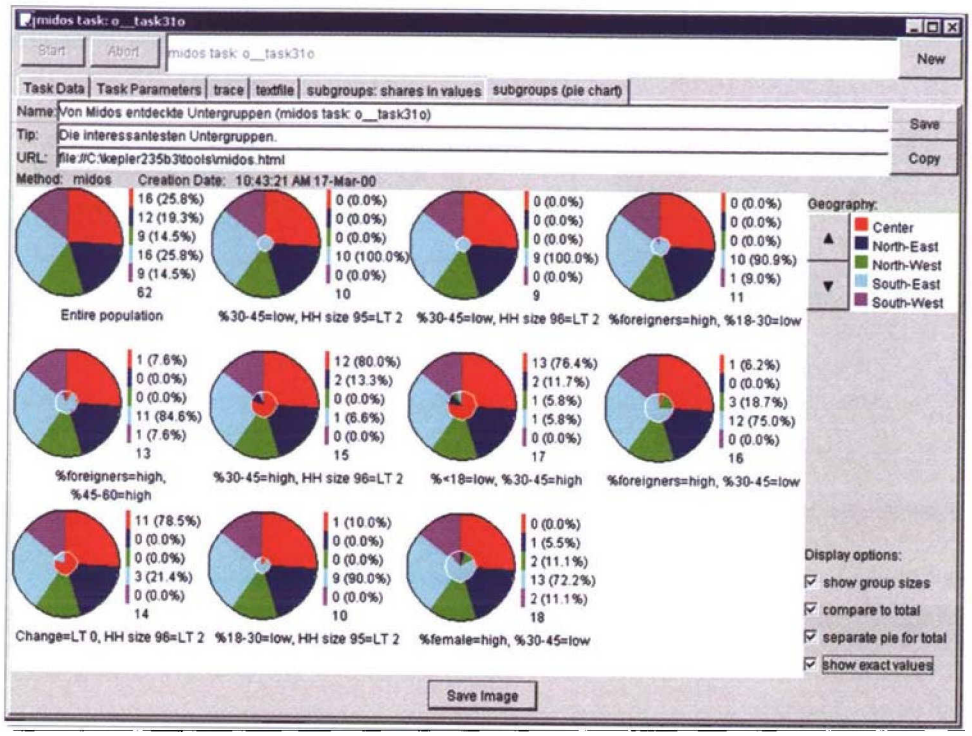

Figure 7. Results of the subgroup discovery method. The groups have been defined on the basis of discretized values of thematic attributes. The pie charts and the numbers on the right of them show proportions of districts from different geographical zones in the whole dataset and in the subgroups.

of the display to the right and from the top to the bottom and skipping the first pie representing the whole set of districts);

- Household size (in 1995 or in 1996) less than or equal to 2 (subgroups 1, 2, and 9);
- Low proportion of the 18-30 age group (subgroups 3 and 9);
- High proportion of foreigners (subgroups 3, 4, and 7);
- High proportion of the 45-60 age group (subgroup 4);
- High proportion of females (subgroup 10).

We already knew about the low percentage of the 18-30 age group in the south–eastern part of the city, as well as about the high percentage of foreigners and of the 45-60 age group, from the application of the previous two data mining methods. However, the other facts are new and provide important additional information for the characterization of this part. Similarly, we learn that most of the central districts are characterized by a high percentage of people between 30 and 45 years old, a low percentage of children up to 18 years old, and a household size of no more than 2 persons (subgroups 5, 6, and 8).

The link between the displays of Descartes and Kepler allows us to investigate the spatial distribution of districts of any subgroup. Let us consider, for instance, the last subgroup of districts in the display in Figure 7 described with the expression "% female=high, % 30-45 = low" (proportion of females in population is high and proportion of people aged between 30 and 45 years is low). This is the largest subgroup found. It contains the highest number of south-eastern districts (13 of 16), and it has characteristics we did not know before. So, we click on the pie corresponding to the group and receive an additional window from Kepler with details about the group (see the left part of Figure 8). Simultaneously, the districts included in the group become highlighted on the map in Descartes shown on the right of Figure 8. On the map we observe that, except for two standalone districts in the north-west, the specified properties refer to a coherent belt in the south and south-east of the city. The belt is formed by almost all districts of the south-eastern part of Bonn and several districts of two other parts adjacent to them.

Thus, with the use of the subgroup discovery method, we were able to gain new knowledge about the center and the south-east of the city. Additionally, we noticed that, probably, there is a spatial trend in proportions of female population and that of people aged between 30 and 45 years. Proportions of women seem to increase from north to south while proportions of people 30 to 45 years old appear to decrease. We let Descartes produce maps visualizing the spatial distribution of these attributes and check whether such a trend really exists. For example, in the map in Figure 9 the districts are divided into four groups of approximately equal size according to the values of the attribute "percentage of female population." It is visible that, indeed, the districts in the south have, in general, higher percentages of female than the

districts in the north. However, there is no strong trend, no continuous increase of values from the north to the south. Instead, one can observe interesting areas of concentration of highest values near the center of the city and on the south–east.

In this section we have demonstrated how an analyst can interactively classify spatial objects according to their locations (and, potentially, any spatial properties). Through such a classification spatial information is encoded in a form suitable

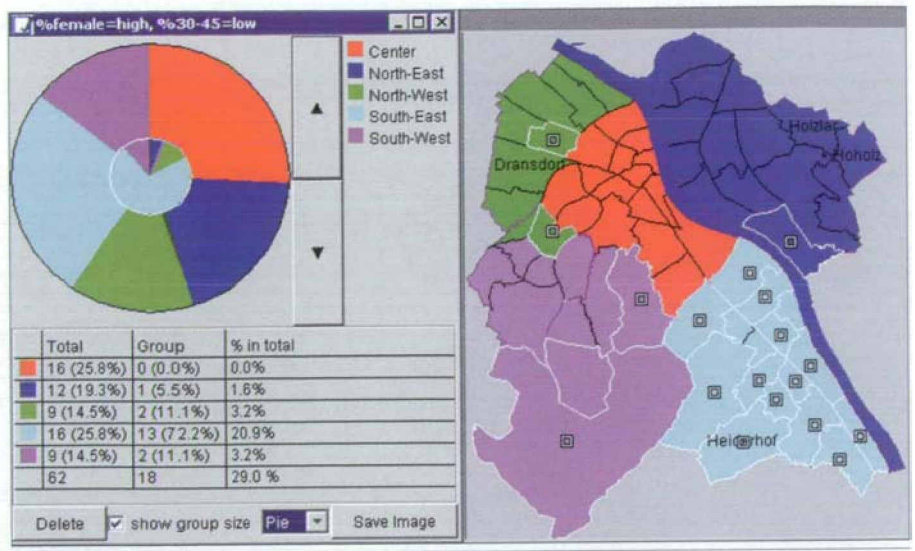

Figure 8. Analysis of spatial distribution of objects included in a selected subgroup.

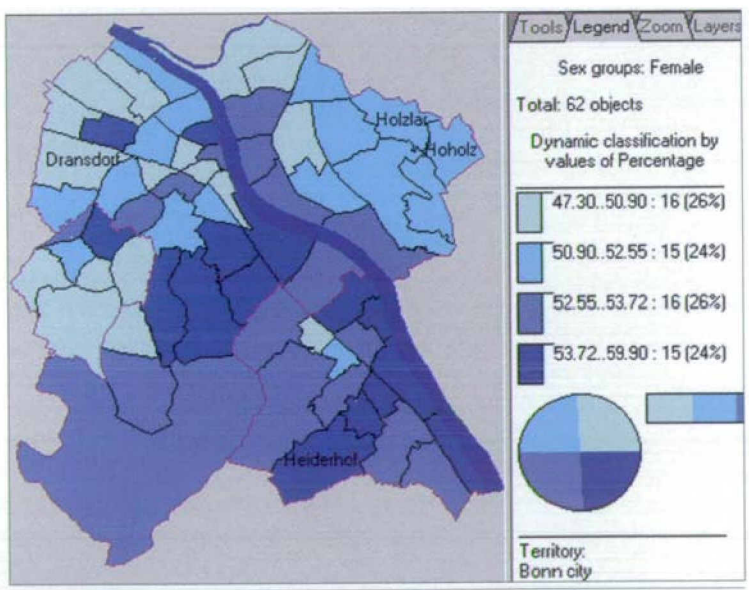

Figure 9. Spatial patterns and trends in the distribution of percentages of female population.

for the application of various data mining methods. These methods may help to relate the spatial properties encoded by means of the classification to thematic characteristics of the spatial objects.

Example 2: Investigating Relationships Between Thematic Attributes

Classification of objects through direct selection of them on a map is just one of many possible methods of defining object classes. Descartes offers users various tools for the classification of objects according to values of one or more thematic attributes (Andrienko and Andrienko 1999a; Andrienko and Andrienko 2001). In this section we present as an example classification utilizing the values of a single numeric attribute. Such classification is done by means of splitting the value range of the attribute into subintervals. Objects with attribute values fitting in the same subinterval compose a class. They have similar appearance on a map (painted in the same color). When these objects are geographical neighbors, they tend to be visually associated into clusters. A classification is "good" from a geographical viewpoint when it produces interpretable, coherent regions on the map. To help an analyst to produce geographically meaningful classes, the classification tool needs to be linked to a map display that shows spatial distribution of the classes. Such a link is provided in Descartes (Figure 10).

A user may break the range of values of a numeric attribute into a desired number of intervals and change the interval boundaries by direct manipulation with the mouse. All changes are immediately reflected in the map display. The map is dynamically repainted as the user moves an interval break. This provides an opportunity to observe geographical properties of many divisions in a short period of time and increases the probability of finding variants producing clear spatial patterns. The user can also pay attention to the statistical distribution of attribute values displayed by a dot plot (see the upper section of Figure 10).

In our experiment, we classified the districts of Bonn according to values of the attribute "House-

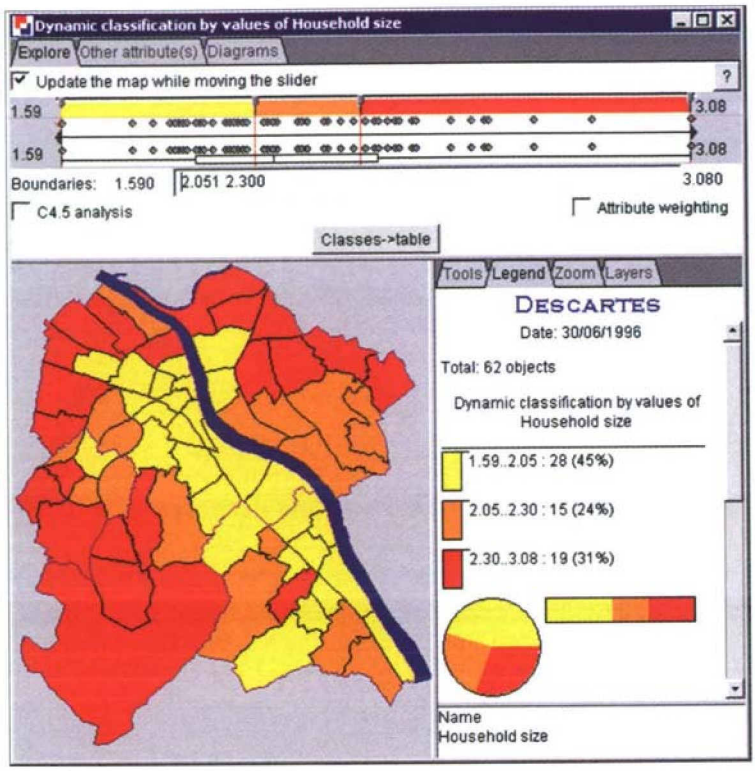

Figure 10. Classification of districts in Descartes by values of household size. The range of values of the attribute is split into three intervals in such a way that the resulting classes of objects form coherent spatial regions.

in terms of percentage of children in total population and percentage of foreigners. The household size is higher where there are more children, which is not surprising, and lower where there are more foreigners, which is an unexpected fact. We have an opportunity to verify this latter fact by using Descartes. The attributes "household size" and "percentage of foreigners" can be visualized on a cross-classification map and a scatter plot (Figure 12). Each district is represented on the scatter plot by a dot whose position is determined by the values of the two attributes considered. The background of the

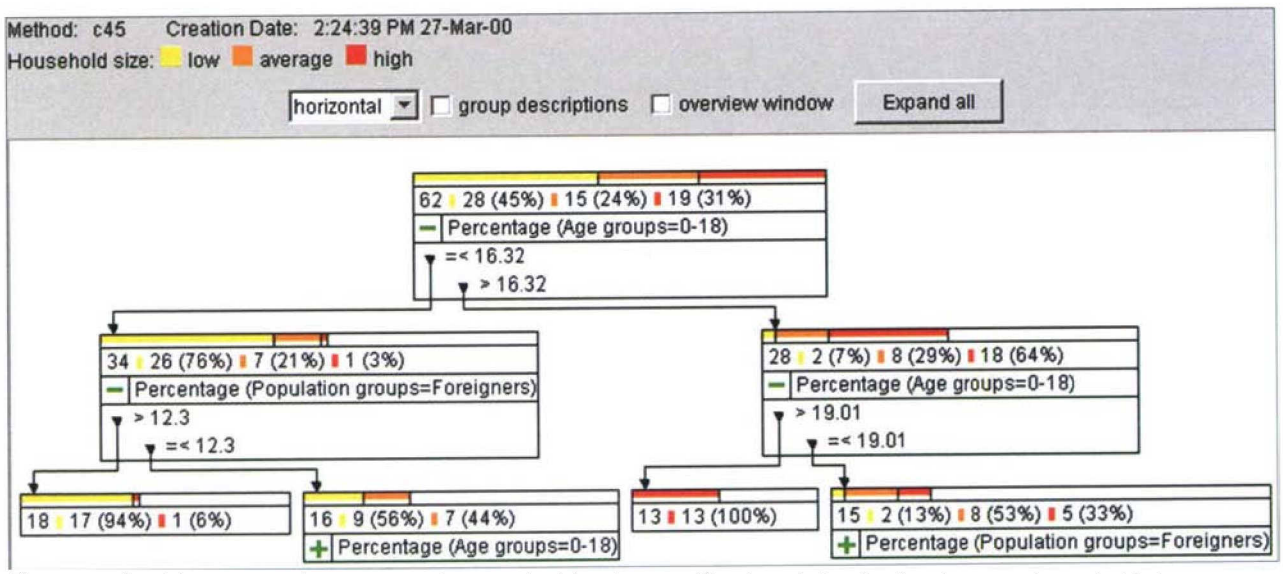

Figure 11. Decision tree produced by the C4.5 method for the classification of city districts by mean household size.

hold size; Date: 30/06/1996" (mean household size on June 30, 1996). We manipulated the number of intervals and their boundaries until coherent geographical areas comprised of neighboring districts with the same colors appeared on the map (Figure 10). Then we passed the classes to the C4.5 algorithm and received the decision tree presented in Figure 11.

From the tree we learned that the classes defined by mean household size significantly differ

scatter plot is divided into differently colored rectangular areas. The colors of the areas are used to paint districts on the map, i.e. each district is painted in the color of the area it fits in. A user of Descartes can divide the background of a scatter plot arbitrarily. In our example, the breaks between the areas correspond to the mean values of the respective attributes.

After removing two outliers from the scatter plot, the resulting distribution of dots shows that

there is a weak inverse correlation between the attributes.[4] Visualization in Descartes has confirmed the findings of the C4.5 algorithm.

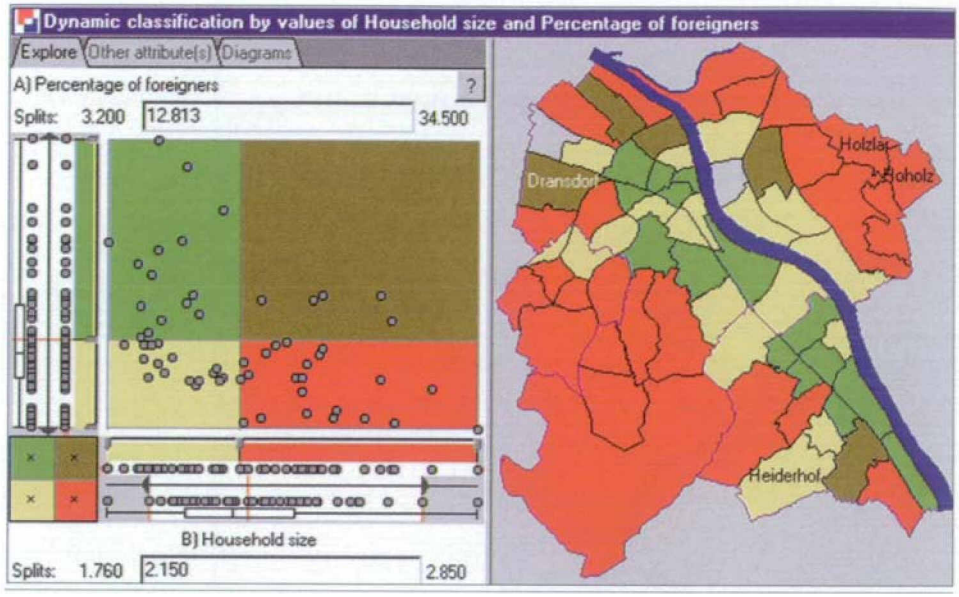

Figure 12. The scatter plot demonstrates a weak inverse correlation between percentage of foreigners and mean household size. The map linked to the scatter plot shows an apparent spatial pattern formed by districts with low number of foreigners and high household size (red districts).

Discussion and Conclusion

The examples described above have demonstrated that interactive visualization and methods of data mining can act as complementary instruments of data analysis. Their integration supports the iterative process of exploratory data analysis that is schematically represented in Table 1. Visualization is used for initial data preview as well as for encoding spatial information in a symbolic form suitable for data mining (e.g., in the form of classes of geographical objects). After that data mining techniques are applied with the aim to reveal previously unknown regularities and relationships in the data set, in particular, to relate spatial and non-spatial characteristics of geographical objects. The results of data mining are interpreted with the help of visual displays that provide the missing spatial reference (e.g., show locations of objects of a group discovered in the process of data mining) or allow the analyst to verify the findings (e.g., check attributes for correlation using a scatter plot). However, it is naïve to expect that a single run of one of the existing data mining methods will discover all the important facts about a given set of data. Therefore, after viewing and verifying the data mining results, it is useful to try other data mining methods, or change the parameters of the previously used method, or modify the input data passed to the method (e.g., change the object classes or select a different subset of thematic attributes). This recurrent procedure corresponds to the nature of the process of data exploration, i.e. search in multiple directions without knowing the outcomes.

Note that in our examples, we applied the system to relatively small data sets. With larger data sets, the synergy of the two approaches is expected to be even more advantageous.

ACKNOWLEDGMENTS

The work on integration of interactive maps and data mining techniques was partly supported within the EU-funded project SPIN! (Spatial Mining for Data of Public Interest, IST Program, project No. IST-1999-10536, January 2000 – December 2002). We are especially grateful to Prof. Terry Slocum for his great interest in our work and valuable comments on and suggestions for improving this paper.

References

Andrienko, G., and N. Andrienko. 1997. Intelligent cartographic visualization for supporting data exploration in the IRIS system. *Programming and Computer Software* 23(5): 268-82.

Andrienko, G., and N. Andrienko. 1999a. Interactive maps for visual data exploration. *International Journal of Geographical Information Science* 13(4): 355-74.

Andrienko, G., and N. Andrienko. 1999b. Knowledge-based visualization to support spatial data mining, in hand, In: Hand, D.J., Kok, J.N., and M.R. Berthold (eds), *Advances in intelligent data analysis.* Proceedings of the 3rd International Symposium, IDA-99, Amsterdam. Lecture Notes in Computer Science 1642, Berlin, Germany: Springer-Verlag. pp.149-60.

Andrienko, G., and N. Andrienko. 1999c. Data mining with C4.5 and cartographic visualization. In: N.W.Paton, and T.Griffiths (eds), *User interfaces to data intensive sys-*

[4] See Andrienko and Andrienko (1999a) for the description of the direct manipulation facility for outlier removal.

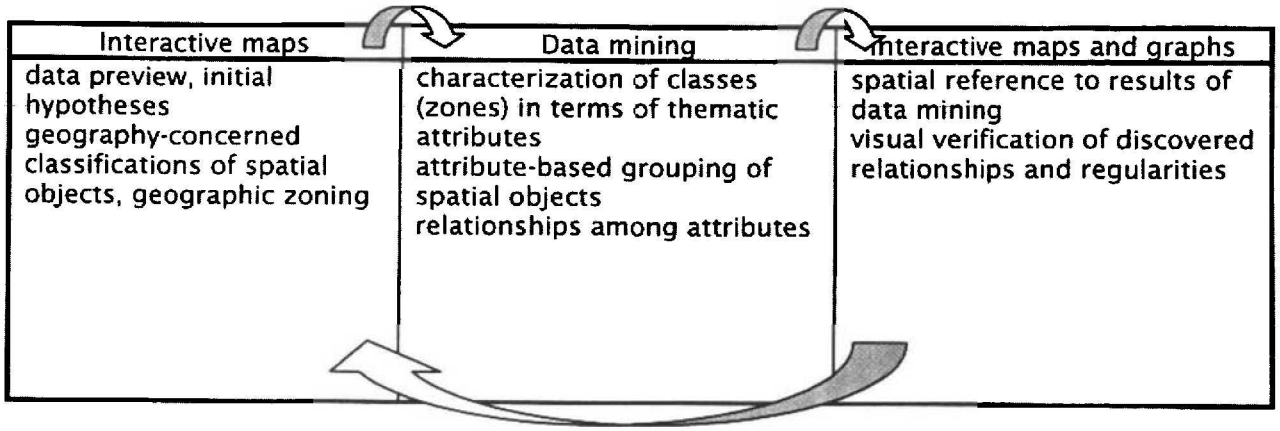

Table 1. Iterative process of exploratory data analysis using interactive visual displays and data mining methods.

tems. IEEE Computer Society, Los Alamitos, California. pp.162-5.

Andrienko, G., and N. Andrienko, H. Voss, H., and J. Carter. 1999. Internet mapping for dissemination of statistical information. *Computers, Environment and Urban Systems* 23(6): 425-41.

Andrienko, G., and N. Andrienko. 2001. Exploring spatial data with dominant attribute map and parallel coordinates. *Computers, Environment and Urban Systems* 25(1): 5-15.

Buja, A., J. A. McDonald, J. Michalak, and W. Stuetzle. 1991. Interactive data visualization using focusing and linking. In: *Proceedings IEEE Visualization'91*. Los Alamitos, California: IEEE Computer Society Press. pp.156-63.

Cook, D., J. Symanzik, J. J. Majure, and N. Cressie. 1997. Dynamic graphics in a GIS: More examples using linked software. *Computers & Geosciences* 23(4): 371-85.

DiBiase, D. 1990. Visualization in the earth sciences. *Earth and Mineral Sciences*, Bulletin of the College of Earth and Mineral Sciences, Penn State University. Vol. 59, pp. 13-18.

Dykes, J.A. 1997. Exploring spatial data representation with dynamic graphics. *Computers & Geosciences* 23(4): 345-70.

Egbert, S.L. and T. A. Slocum. 1992. EXPLOREMAP: An exploration system for choropleth maps. *Annals of the Association of American Geographers* 82: 275-88.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. 1996. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM* 39 (11): 27-34.

Gebhardt, F. 1997. Finding spatial clusters. In: *Principles of data mining and knowledge discovery*. Berlin, Germany: Springer-Verlag. pp. 277-87.

Guittet, C., M. van Liedekerke, P. Loekkemyhr, N. Andrienko, G. Andrienko, P. Gatalsky, H. Voss, and D. Schmidt. 2001. EuroFigures—A digital portrait of the EU's general statistics. In: Pre-proceedings NTTS&ETK, Eurostat. Vol. 2, pp. 937-8.

Klösgen, W. 1996. Explora: A multipattern and multistrategy discovery assistant, In: U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds), *Advances in knowledge discovery and data mining*. Cambridge, Massachusetts: MIT Press. pp. 249-71.

Koperski, K., J. Han, and N. Stefanovic. 1998. An efficient two-step method for classification of spatial data. In: *Proceedings SDH'98*, Vancouver, Canada. International Geographical Union, pp.45-54.

McDonald, J.A. 1982. *Interactive graphics for data analysis*. Project Orion, 11, Stanford, California: Stanford University,

MacDougall, E.B. 1992. Exploratory analysis, dynamic statistical visualization, and geographic information systems. *Cartography and Geographic Information Systems* 19 (4): 237-46.

MacEachren, A. M. 1994. Visualization in modern cartography: Setting the agenda. In: *Visualisation in modern cartography*. New York, New York: Elsevier Science Inc. pp.1-12.

MacEachren, A.M., and M.-J. Kraak. 1997. Exploratory cartographic visualization: Advancing the agenda. *Computers and Geosciences* 23(4): 335-44.

Openshaw, S., M. Charlton, C. Wymer, and A. Craft. 1987. A Mark I geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographic Information Systems* 1: 335-58.

Quinlan, J.R. 1993. *C4.5: Programs for machine learning*. San Mateo, California: Morgan Kaufmann Publishers.

Symanzik, J., J. Majure, and D. Cook. 1996. Dynamic graphics in a GIS: A bidirectional link between ArcView 2.0 and XGobi. *Computing Science & Statistics* 27: 299-303.

Tukey, J.W. 1977. *Exploratory data analysis*. Reading, Massachusetts: Addison-Wesley.

Wettschereck, D., D. W. Aha, and T. Mohri. 1997. A review and empirical evaluation of feature weighting methods for a class of Lazy Learning algorithms. *Artificial Intelligence Review* 11: 273-314.

Wrobel, S. 1997. An algorithm for multi-relational discovery of subgroups. In: J.Komorowski, and J. Zytkow (eds), *Principles of data mining and knowledge discovery*. Lecture Notes Computer Science 1263. Berlin, Germany: Springer-Verlag. pp. 78-87.

Wrobel, S., D. Wettschereck, E. Sommer, and W. Emde. 1996. Extensibility in data mining systems. In: *Proceedings of KDD'96 2nd International Conference on Knowledge Discovery and Data Mining*. Menlo Park, California: AAAI Press, pp.214-19.

Zhang, Z., and D. A. Griffith. 1997. Developing user-friendly spatial statistical analysis module for GIS: An example using ArcView. *Computers, Environment and Urban Systems* 21(1): 5-29.