IJPIS

KNOWLEDGE DISCOVERY IN ROAD ACCIDENTS DATABASE

- INTEGRATION OF VISUAL AND AUTOMATIC DATA MINING METHODS

IMAD ABUGESSAISA
Department of Computer and Information Science
Linköping University
SE-581 83 Linköping
Sweden
g-imaab@ida.liu.se

Abstract

Road accident statistics are collected and used by a large number of users and this can result in a huge volume of data which requires to be explored in order to ascertain the hidden knowledge. Potential knowledge may be hidden because of the accumulation of data, which limits the exploration task for the road safety expert and, hence, reduces the utilization of the database. In order to assist in solving these problems, this paper explores Automatic and Visual Data Mining (VDM) methods. The main purpose is to study VDM methods and their applicability to knowledge discovery in a road accident databases. The basic feature of VDM is to involve the user in the exploration process. VDM uses direct interactive methods to allow the user to obtain an insight into and recognize different patterns in the dataset. In this paper, I apply a range of methods and techniques, including a paradigm for VDM, exploratory data analysis, and clustering methods, such as K-means algorithms, hierarchical agglomerative clustering (HAC), classification trees, and self-organized-maps (SOM). These methods assist in integrating VDM with automatic data mining algorithms. Open source VDM tools offering visualization techniques were used. The first contribution of this paper lies in the area of discovering clusters and different relationships (such as the relationship between socioeconomic indicators and fatalities, traffic risk and population, personal risk and car per capita, etc.) in the road safety database. The methods used were very useful and valuable for detecting clusters of countries that share similar traffic situations. The second contribution was the exploratory data analysis where the user can explore the contents and the structure of the data set at an early stage of the analysis. This is supported by the filtering components of VDM. This assists expert users with a strong background in traffic safety analysis to be able to intimate assumptions and hypotheses concerning future situations. The third contribution involved interactive explorations based on brushing and linking methods; this novel approach assists both the experienced and inexperienced users to detect and recognize interesting patterns in the available database. The results obtained showed that this approach offers a better understanding of the contents of road safety databases, with respect to current statistical techniques and approaches used for analyzing road safety situations.

Keywords: Visual data mining, K-Means, HAC, SOM , InfoVis, IRTAD, GLOBESAFE

1. Introduction to Data Mining (DM)

Digital data acquisition methods and storage technology have resulted in the growth of a huge amounts of data being stored in different types of databases. With this

advancement in database technologies, the need to extract useful information from the database is increased. The field concerned with this task has become known as *data mining* [Pang-Ning et al., 2006].

"*Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and summarize the data in novel ways that are both understandable and useful to the data owner*" [Hand et al., 2001]. A more relevant definition is given by Demšar [2006] "*Data mining is the process of identifying useful and as yet undiscovered structure in the database*".

The relationships and summaries derived through data mining process are models, patterns or relationships. The structure found within a set of the data must be novel, and the novelty must be measured relative to the user's prior knowledge. The structure should also be understandable. This implies that simplicity of structure is required to make the results clear and understandable.

1.1. Knowledge Discovery in Databases (KDD)

Knowledge discovery in databases originated in the field of Artificial Intelligence (AI); and DM has been placed within the context of KDD. As a process, KDD involves several stages. The process starts by selecting the target data set, transforming it to the required format using data mining software (optional stage), performing data mining to extract the structure, and finally, interpreting and assessing the discovered structure [Hoffman, 1997].

Many steps precede the data mining steps: retrieving the data from large databases, selecting the appropriate subset to work with, deciding upon the appropriate sampling strategy, cleaning the data and dealing with missing fields, applying appropriate transformations, dimensionality reductions, and projections. To decide whether this extracted information does represent knowledge, the information is evaluated by visualizing it. The stage of visualization and visual interaction between the computer and the user is known as visual data mining (VDM) [Keim, 2002]. Finally, the knowledge should be consolidated with the existing knowledge.

1.2. Applications of Data Mining

DM has been used in vast array of areas; this section discusses and surveys data mining applications in business and science. In the area of banking, DM has been used for financial data analysis. The data collected in the banking and financial industries are often relatively complete, reliable and of high quality. These features facilitate systematic data analysis and DM. One way DM analysis is used in banking is to classify and cluster customers for targeted marketing. Classification and clustering methods can be used for group identification of customers and their targeted marketing. An example of this classification is to identify the most crucial factors that influence a customer's decisions regarding banking. Customers with similar behaviors regarding loan payments, for example, may be identified by multidimensional clustering techniques.

Another application of DM in banking and the finance industry is the detection of laundering and other financial crimes. In this case, different information is integrated from multiple databases related to the study. Analysis tools can be used to detect unusual patterns, such as significant cash flows during particular periods of time by a certain group of customers.

The telecommunications industry is a growing industry. The development of this sector has created a great demand for data mining in order to help understand the

business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of services.

Within the scientific field, DM has been applied to biological analysis. Biological data mining has become an essential part of a new research field called bioinformatics. The identification of DNA and amino acid sequence patterns play important roles in bioinformatics research and genetic engineering.

2. Visual Data Mining (VDM)

As discussed above, DM and KDD assist in the exploration of hidden and undiscovered structures in the database. Algorithmic and automatic DM make use of the computational capabilities of computer hardware. On the other hand, humans with their perceptual capabilities are not involved in the exploration process. The general approach involving humans in the process of DM is VDM, "*VDM aims at integrating the human in the data exploration process, applying its perceptual abilities to the large data sets available in today's computer systems*" [Keim, 2002].

VDM presents the data in a visual form that assists humans to interact with the data and draw conclusions from it. The VDM process is based on the conditions that: (1) humans have little knowledge about the data; (2) the exploration task is vague.

2.1. Paradigm for VDM

The paradigm for VDM is described in Stuart [1999]. Figure 1 (below) shows the steps involved in this paradigm.

The user starts by obtaining an overview of the data context. This step helps to identify interesting areas (patterns) in the data, and then focuses on specific patterns. After obtaining an overview of the data, the user can then start to analyze the selected patterns, in step 2. This is performed by accessing the details of the data.

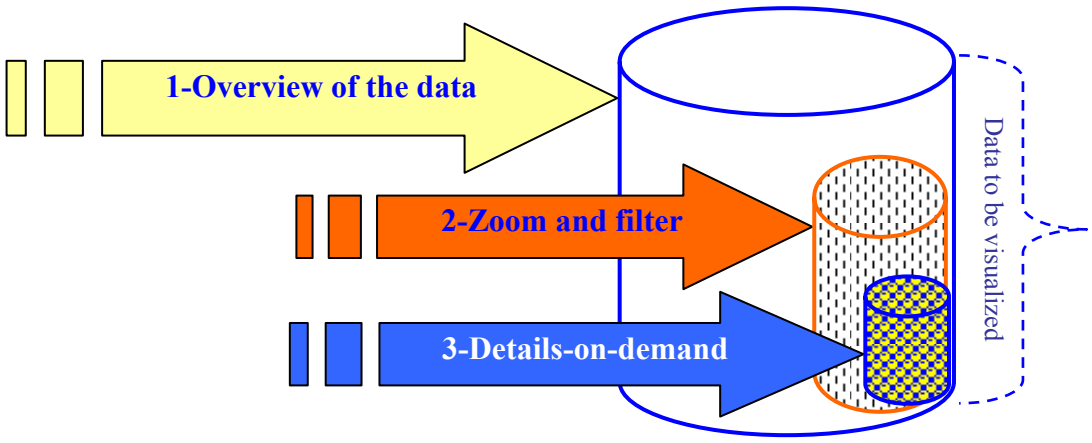

Figure 1. Paradigm for VDM.

2.2. Classification of VDM Techniques

Information Visualization (InfoVis) focuses on the mapping of abstract data onto the screen space [Shneiderman, 1996]. Many techniques have been developed, some of which have good exploration abilities, but can only be used for low dimensional data sets. The requirement for visualization techniques that assist in the visualization

process for multi-dimension data sets is continually increasing and thus many tools have been developed for this purpose. These techniques can be classified according to three aspects, see Table 1 [Keim, 2002]:

- The data to be visualized,
- The visualization technique
- The interaction and distortion technique used

Data type to be visualized	Visualization technique	Interaction and distortion techniques
One-dimensional data	Standard 2D/3D displays	Interactive projection
Two-dimensional data	Geometrically transformed displays	Interactive filtering
Multidimensional data	Icon-based displays	Interactive zooming
Text and hypertext	Dense pixel displays	Interactive distortion
Hierarchies and graphs	Stacked displays	Interactive linking and brushing

Table 1. Classification of VDM techniques.

2.3. Visualization Methods

Knowledge discovery and mining tasks can be easily performed by means of graphical presentations and this is the role of visualization [Fayyad, 2002]. The main function of visualization is to: (1) Provide an overview that simplifies complex data sets; (2) Summarize the data and help to identify the relationship and patterns in the data set [Breunig, 2001].

InfoVis principles can be applied in order to visualize the results obtained by data mining algorithms. Methods such as scatterplots [Harris, 1999], Radviz, and histograms provide information regarding the relationship between two numerical attributes and a discrete one. All are applied as a geometric visualization method where certain value(s) are visualized as points in a two-dimensional space and the values of attributes only influence the position of the point and not its size, shape or color [Grinstein, 2001]. In this research those methods were implemented. Furthermore, for the visualization of massive data sets, graphical visualization methods were used [Tufte, 1986]. These are discussed below.

2.3.1. Histograms

Histograms estimate the probability distribution for a certain numerical attribute of a given set of objects with a single dimension [Ioannidis and Christodoulakis, 1993]. The histogram represents the number of instances in the data set for a particular attribute (field). The main advantage of the histogram method, with respect to other methods, is its ability to offer a description of a large dataset. It is a popular form of data reduction, and it is highly effective at approximating both sparse and dense data sets. Figure 2 shows an example of a histogram for cars per capita in the Organization for Economic Co-operation and Development (OECD) (source of data IRTAD) countries in 1970.

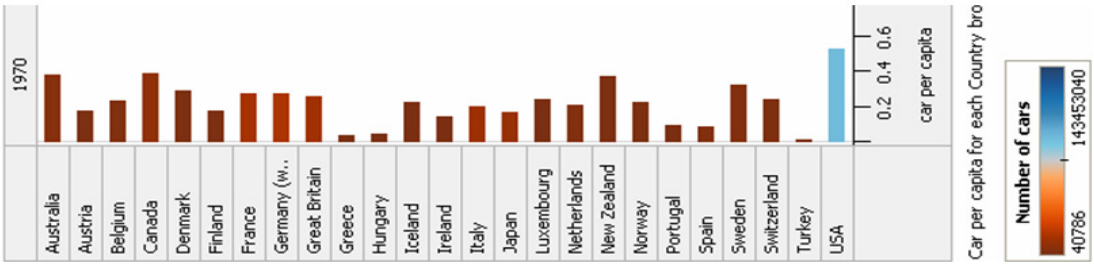

Figure 2. Histogram of cars per capita.

2.3.2. Dendrogram

A dendrogram is a tree structure used to represent the process of hierarchical clustering. The tree shows how objects (countries) are grouped together step by step for visualization by HAC (A clustering method that produces "natural" groups of examples, characterized by attributes, see section 5.3.1.)

The means by which a dendrogram is interpreted is illustrated in Figure 3, which shows a dendrogram for five objects (m = 0, m = 1, m = 2, m = 3, and m= 4). m = 0 shows the five objects as singleton clusters at level 0, objects **a** and **b** are grouped together to show the first cluster, and stay together in all the subsequent levels. The vertical axis shows the similarities between clusters and in this case the similarity of two groups of objects {**a**, **b**} and {**c**, **d**, **e**} is 0.1. These are merged together to form a single cluster {**a**, **b**, **c**, **d**, **e**}.

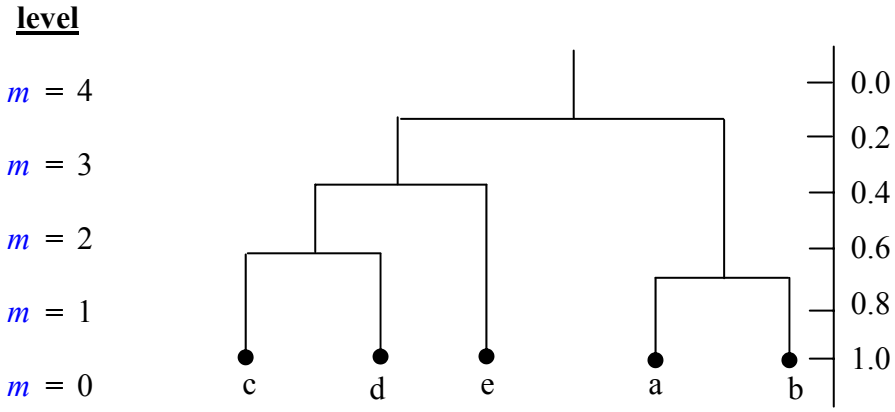

Figure 3. An illustration showing the manner in which a dendrogram presents the clusters in levels from (m_0...m_4) and a similarity scale (1.0...0.0).

2.3.3. Parallel Coordinates (PC)

The PC method is used in VDM to represent multi-dimensional (m-dimensional) information in 2 dimensional space [Inselberg, 1981]. With parallel coordinates, all m-axes are vertical, parallel to one another and equally spaced. The axes are linearly scaled and their values are arranged from the minimum to maximum according to the value of the attributes in the data set. Each item from the data set is presented as a polygonal line intersecting each of the axes at the point that corresponds to the data value.

PC can also be used to illustrate the value of one item (to be represented as the target item) and visualize its difference to other data items. Linking and brushing can be used to highlight the characteristics of the data such as data distributions and

clusters. Figure (4) shows a PC with 5 m (m1...m5) and scale from i-j and within the same figure, two items are represented by red and green polygonal.

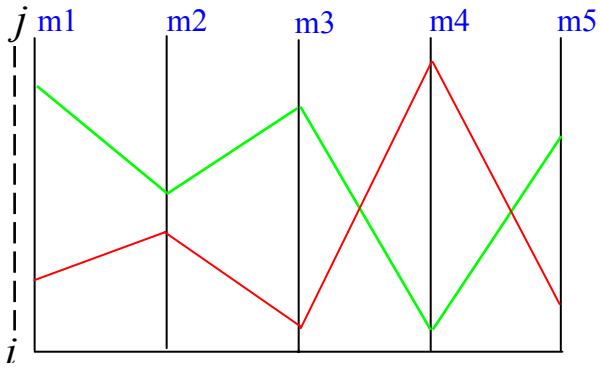

Figure 4. Parallel coordinates with five dimensions (axes) (m1…m5) each corresponding to particular attributes and in which the axes are linearly scaled. Their values are arranged from the minimum to maximum (i...j).

2.3.4. Radviz

Radviz stands for radial visualization. This visualization method is based on mapping *m*-dimensional points onto a two dimensional space [Hoffman et al., 1997]. The *m* points (feature or data set attributes) are equally spaced around the circumference of a circle as shown in Figure 5.

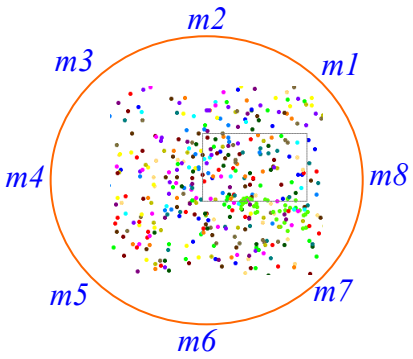

Figure 5. Radviz with eight feature points equally spaced around the circumference of the circle (m1…m8).

The data set instances are plotted as colored points inside the circle and each of the colored points is located in position by springs that are attached at the other end to the attribute anchor point.

3. Previous Work

This research has been on-going since 2004. The main objective is to make road traffic and road accident information and knowledge easily shared and exchanged between road safety experts. The previous work based on a conceptual model with three layers has been implemented as a platform, GLOBESAFE (http://www.globesafe.org/). The components of the model and the relationships between them are described in [Abugessaisa et al., 2007]. The research investigated different methods and applications for road safety analysis and benchmarking.

4. Research Problem

Road safety experts and researchers deal with large volumes of quantitative information and collected statistics, in order to understand and estimate the social and economic cost of the accidents and to be able to introduce safety plans in order to prevent or reduce occurrences of accidents. The road traffic and accident statistics must be presented in such a way to make it easier to be both recognized and interpreted by a human operator.

Previous work on accident analysis included statistical methods and formal techniques [Gitelman and Hakkert, 1997; Gurubhagavatula, 2008]. Statistics tables and ordinary charting techniques are not sufficient for present day requirements and this causes difficulties in the effective visualization of results and patterns. Another disadvantage is that ordinary methods limit human involvement in the exploration task. The dimensions of the data have a strong influence on the understanding and exploration tasks, the higher the dimensions of the data, the more difficult the exploration and visualization task becomes [Theus, 2005].

This paper introduces a novel approach to knowledge discovery in road databases using visual data mining methods.

5. Tools, Methods and Materials

This section of the paper discusses the methods used and describes the software tools used to apply VDM methods and clustering algorithms. A dataset extracted from two road safety databases was analyzed.

5.1. International Road Accident Databases and the Data Set Used

In developing an accident database at the national level, efforts were made to use the technologies available that are able to assist in developing and implementing road safety programs and enable planning at both the local and national levels. The study of these databases showed that a tremendous effort had been made by different international and regional organizations[1]. The main objectives of these databases were to explain road accidents in compatible and homogenous formats and to reduce the effort and time spent by researchers and end users in collecting accident statistics. International and regional road safety organizations use these databases to publish annual reports and statistics, according to predefined user requirements with agreed variables and indicators.

Research in this area is devoted to developing database schema that represent the road accident data and other relevant statistics (population censuses, socioeconomic and energy related data) that is useful for users and decision-makers. Technically, all available road databases collect and process data in different ways [Jaroslav and Mikulik, 2005].

Most of the accident variables used in different databases are selected to give full descriptions of the crash type and persons involved [Yannis el al., 1998]. The output provides users with the results after manipulating the database. The results of the manipulation could be represented using different charting techniques in order to make the interpretation and investigation more understandable by using visualization techniques.

[1] WHO, EU and IRTAD.

For the purpose of this research, the data set used comes from two international road databases that have good quality information and have been in existence for a significant length of time.

5.1.1. IRTAD (http://www.irtad.net/)

The International Road Traffic and Accident Database (IRTAD) is an international database that gathers data on traffic and road accidents (accident and victims as well as exposure data are collected on a continuous basis) from the Organization for Economic Co-operation and Development (OECD) member countries. The main part of the database, with around 500 data items, includes aggregated data on injury accidents, road fatalities, vehicle population, and network length. The data is collected from 28 countries (for 1965 and for every year since 1970). The data set was offered by IRTAD under a special agreement with the author. The content of the database is represented in Figure 6.

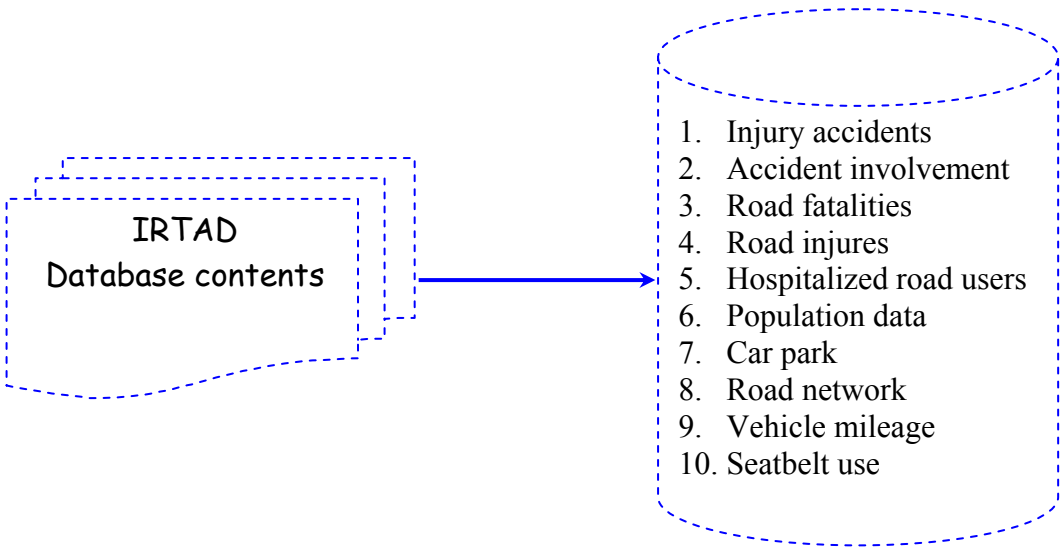

Figure 6. Content of IRTAD database aggregated by country and year for OECD members.

5.1.2. GLOBESAFE (http://www.globesafe.org/)

GLOBESAFE is a database and a platform used to share road safety information among road safety organizations. This database offered data for the nine ASEAN countries from 1994-2002 [Abugessaisa et al., 2007]. The content of the database is shown in Figure 7.

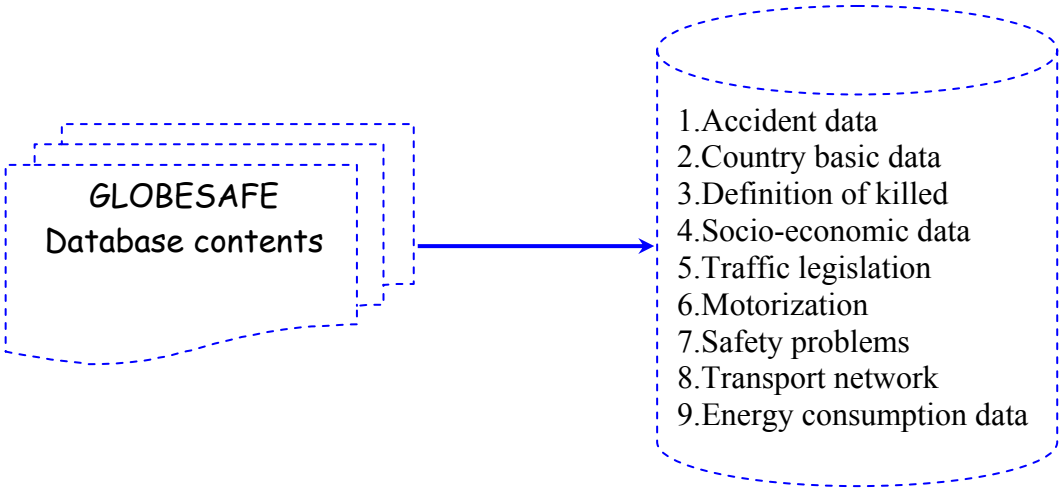

Figure 7. Content of GLOBESAFE, database aggregated by country and year for countries from developed and developing world.

5.2. Tools

To perform machine learning and apply DM algorithms, two tools were used; the first was Orange, which is machine learning and data mining software [Demsar et al., 2004]. Orange is a comprehensive, component-based framework for both experienced data mining and machine learning. Among other features, Orange has the ability to accept a variety of popular data formats as its input. It also supports different visualization techniques that provided good results in this research.

The second tool used was Tanagra, which is free DM software for academic and research purposes. It supports several data mining methods that include, but are not limited to, Exploratory Data Analysis (EDA), statistical learning, machine learning and limited visualization techniques (dendrogram and scatterplot) [Rakotomalala, 2005].

5.3. Methods

5.3.1. Clustering methods

A clustering problem is an unsupervised learning problem [Soukup and Davidson, 2002]. The aim was to find clusters in the data of similar countries sharing a number of interesting properties (level of motorization, number of fatalities, same network length, etc.). Clustering methods were used to find the similarities among those countries in the same or different regions. Clustering methods help to automatically represent a data set by a small numbers of regions, preserving the topological properties of the original input space.

K-means algorithm

With distance-based clustering methods, the K-means technique is considered to be a classical clustering method [Jain et al., 1999]. One of the main features that make K-means useful in this research is the possibility of specifying in advance how many clusters are being sought. This number refers to the parameter K (here it refers to the number of countries). From this number, K points are chosen in a random cluster as cluster centers. All instances are assigned to their closest cluster center according to the Euclidean distance metric. The next step is to calculate the centroids or the mean of the instances in each cluster. Then these centroids are taken to be the new center

values. The results of K-means algorithms applied to a data set of energy consumption are plotted in scatterplots in Figures 12 and 13. The K-means technique is simple and effective, giving results that are verifiable and useful.

Self-organizing map (SOM)

A special version of K-means, SOM is one of the most popular neural network clustering methods [Han and Kamber, 2006]. The main goal of SOM is to represent all points in a high-dimensional source space by points in a low-dimensional (e.g. 2D) target space. This ensures the distance and proximity relationships in the cluster are preserved. Clustering in SOM is performed by having several units competing for the current object (cell of the map); hence, the units whose weight vector is closest to the current cell becomes the wining or active cell (Figures 21-23).

Hierarchical Agglomerative Clustering 'HAC'

HAC is a clustering method that produces "natural" groups of examples, characterized by attributes. HAC uses a bottom-up strategy, which starts by placing each object in its own cluster and then merging these atomic clusters into larger and larger clusters until all of the countries are in a single cluster [Rakotomalala, 2005]. A dendrogram is commonly used to represent the process of hierarchical clustering (Figure 17). A dendrogram tree represents the successive agglomerations, starting from one country per cluster, until such time as the whole dataset belongs to one cluster. The main advantage of HAC is that the user can guess the correct partitioning by visualizing the tree. A disadvantage is that it requires the computation of distances between each example, which is very time consuming when the dataset is large.

Decision Tree (DT)

A decision tree is a classification tree, which determines an object's class by following the path from the root to a leaf node [Poole, 1998]. It chooses branches according to the attribute value of the objects. A decision tree is induced from the training set. Classification rules can also be extracted from the decision tree.

The DT method uses 'divide-and-conquer', nodes, which involves the testing of a particular attribute. The test at a single node compares an attribute with a constant. Leaf nodes give a classification that applies to all instances that reach the leaf, or a set of classifications, (Figures 10-11).

5.3.2. Building Knowledge Flow

The first step that corresponds to data selection in the data mining process is building the knowledge flow. Knowledge flow represents the knowledge discovery process as a stream diagram [Demsar, 2004]. The main steps in building the knowledge flow are to:

- Import the dataset, i.e. linking the database table; excel sheet, or text format file to the software
- Compute descriptive statistics
- Select target and input attributes (discrete and continuous)
- Choose the learning algorithm
- Build the prediction model and visualize the results

Both Tanagra and Orange support building knowledge flow using a drag-and-drop interface. Figure 9 shows a knowledge flow diagram for computing DT and other algorithms.

5.3.3. Brushing and Linking: Enhanced User Interaction

Visual data mining tools support the use of different visualization methods, as described in Section 2.3. These methods can be used independently for the same data set. Combining different visualization methods to mine and visualize the same data set is made possible by applying linking and brushing techniques (Figure 8).

In VDM there are two different types of interaction that can be offered by the tools used [Demsar, 2006]:

- Single Visualization

This allows the user to explore the content of the visualization during the VDM process and graphically extract a subset of the attributes. Graphical entities in the VDM can represent a particular value of an attribute or a particular instance from the data set. Each of the visualization methods described in Section 2.3. has its own graphical entities, for example, in a histogram, a bar is a graphical entity, the raw information for the data table, the line of the parallel coordinates, a path in the DT (traversing from leaf node), and so forth. This makes it possible for the user to examine a subset of the data object(s), by clicking on the graphical entity or selecting from the set of attributes using a dialogue window.

- Brushing and Linking

The idea of linking and brushing techniques is to combine different visualization methods in order to overcome the shortcomings of a single technique [Keim, 2002]. By combining multiple visualizations, users can obtain more information than from a single method. Another feature of brushing and linking, involves the possibility of highlighting subsets of the data points, or removing and reducing the number and size of the points. The user can apply these techniques to *de-emphasize the subset, if he/she wants to focus on the rest of the set (*Voigt, 2004). With these techniques, the user can add or remove attribute(s) and alter their appearance. Brushing and linking methods and their role in VDM are illustrated in Figure 8.

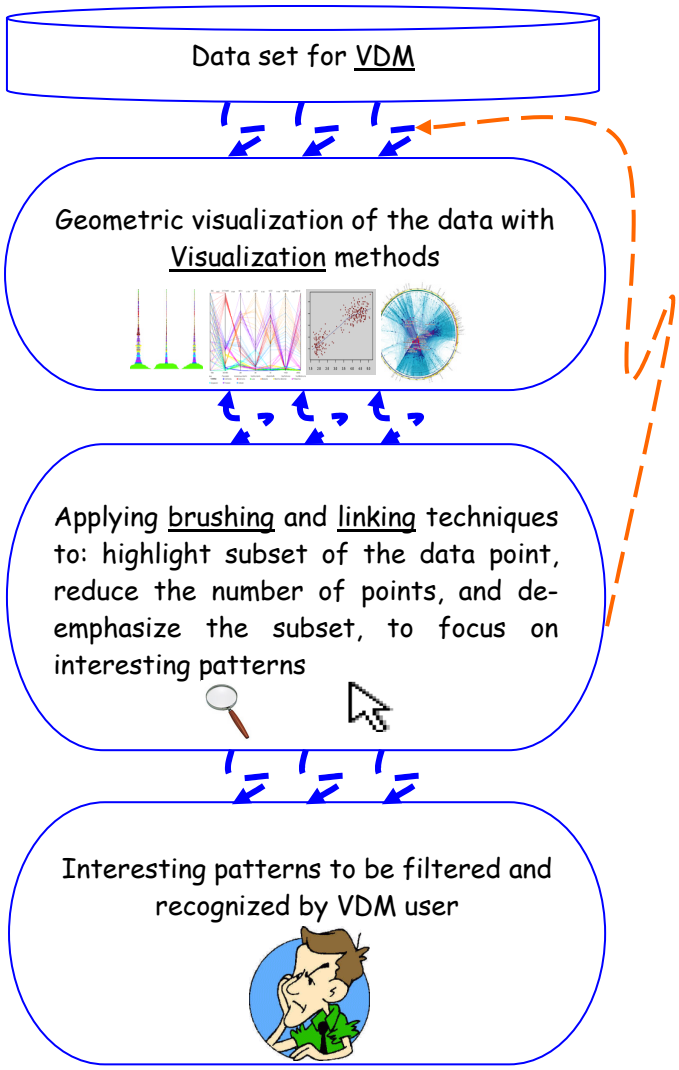

Figure 8. Illustration of brushing and linking techniques in the VDM process from data selection to pattern recognition and filtering.

6. Results

This paper discussed and presented applications of VDM for traffic and accident data. For these applications, different areas were identified in which VDM was able to be combined with automatic algorithms of data mining. The first area involved the discovery of clusters and different relationships (such as the relationship between socioeconomic indicators and fatalities, traffic risk and population, personal risk and car per capita, etc.) in the road safety database for two regions ASEAN and OECD. The clustering methods used are HAC, DT, K-means, and SOM. The results shown in this section suggested that these methods were very useful and valuable for detecting clusters in countries which share similar traffic situations, such as the total number of accidents, the number of fatalities and injuries, the consumption of energy within the transport sector and their CO_2 emissions. The second application was the exploratory data analysis in which the user was able to explore the contents and the structure of the data set at an early stage of the analysis. This is supported by the filtering components of VDM. EDA supports observations and detection of missing and noisy data. Expert users, with a strong background in traffic safety analysis, will be able to determine assumptions and hypotheses concerning future situations.

The third contribution involved interactive explorations based on brushing and linking methods; this novel approach assists both the experienced and inexperienced users to detect and recognize interesting patterns in the available database. This is possible because of the visual ability of the users to detect and recognize patterns (Nichols and Newsome, 1999).

6.1. Knowledge Flow and Classification Tree

Figure 9 shows the knowledge flow for the mining of accident-related statistics from the ASEAN region, through the use of DT techniques. The tree obtained is shown in Figures 10 and 11. The tree shows the majority class probability. The target class in this tree was Malaysia. Each node has a color according to the majority class probability.

A classification tree offers a potentially more comprehensible solution which possesses a high degree of accuracy. For this data set, however, the DT obtained is somewhat smaller than is possible for other data sets, which can be up to a factor of ten larger in some experiments. The implication of this is that the larger data set the larger is the DT.

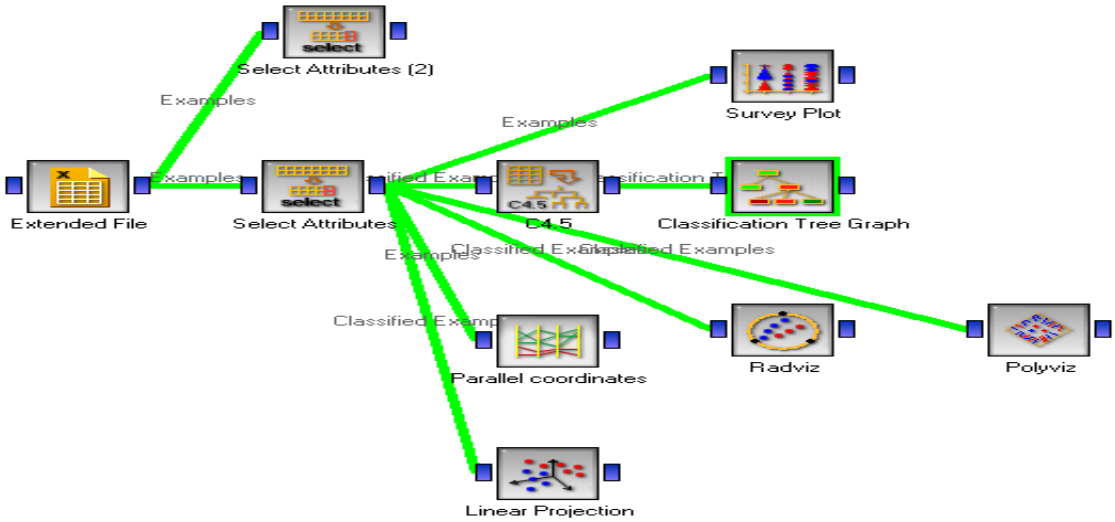

Figure 9. Knowledge flow for mining accident statistics from ASEAN countries.

The DT (Figure 10) makes it possible to compare groups of countries, with respect to their similarities; the group is usually illustrated by a sub tree. The figure shows that that Laos and Cambodia, Myanmar and Malaysia, Indonesia and Thailand, and Philippines and Vietnam are grouped in different sub trees at the lowest level of DT.

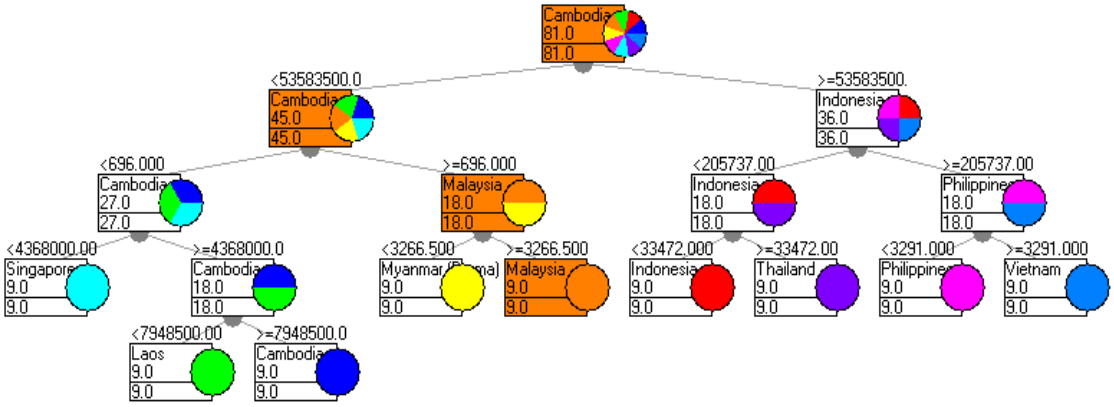

Figure 10. Tree graph, with Malaysia as target variable.

Classification Tree	Class	P(Class)	P(Target)	#Inst
<root>	Cambodia	11	11	81
Population <53583500.000	Cambodia	20	20	45
killedintraffic <696.000	Cambodia	33	33	27
Population <4368000.000	Singapore	100	0	9
Population >=4368000.000	Cambodia	50	50	18
Population <7948500.000	Laos	100	0	9
Population >=7948500.000	Cambodia	100	100	9
killedintraffic >=696.000	Malaysia	50	0	18
killedintraffic <3266.500	Myanmar (Burma)	100	0	9
killedintraffic >=3266.500	Malaysia	100	0	9
Population >=53583500.000	Indonesia	25	0	36
NoofVehicles <205737.000	Indonesia	50	0	18
Injuriesinaccidents <33472.000	Indonesia	100	0	9
Injuriesinaccidents >=33472.000	Thailand	100	0	9
NoofVehicles >=205737.000	Philippines	50	0	18
killedintraffic <3291.000	Philippines	100	0	9
killedintraffic >=3291.000	Vietnam	100	0	9

Figure 11. Classification of the tree graph, with Malaysia as the target variable and offering a range of values for different variables.

6.2. Direct and Inverse Relationship by K-means and HAC

Table 2 shows the results for the K-means clustering, applied to the same set of data. The mining task explored the main clusters that could be found within the data set.

The result for K-means is 5 clusters and a maximum iteration of 10. In this case, 5 trials were performed with distance normalization variance and average computation, as shown in Table 2 the best ratio obtained in trial 2 is equal to 0.877031.

Cluster	Countries	Size	Trial	Ratio explained
cluster n°1	Malaysia	9	1	0.870533
cluster n°2	Thailand	9	2	0.877031
cluster n°3	Cambodia, Laos, Myanmar, Philippines, Singapore	45	3	0.863480
cluster n°4	Indonesia	9	4	0.758948
cluster n°5	Vietnam	9	5	0.775168

Table 2. K-Means clustering results, Number of clusters and trail = 5.

The K-means results are presented by means of the scatterplots in Figures 12 and 13, in which the *x*-axes represent the total number of people killed in traffic and the *y*-axes, the total number of motorcycles. All the data items are plotted in the display area according to their attribute values. The color and shape of the data points in the figures are assigned according to the cluster number. The scatterplot indicates the relationship between numbers of people killed in traffic against the number of motorcycles. According to the scatterplot, *C_Kmeans_2* contains a single country, Thailand (represented by a green cross), which has a high number of people killed in traffic in the ASEAN region. *C_Kmeans_4* (blue cross) has the second highest number of people killed. This cluster contains a single country, Indonesia.

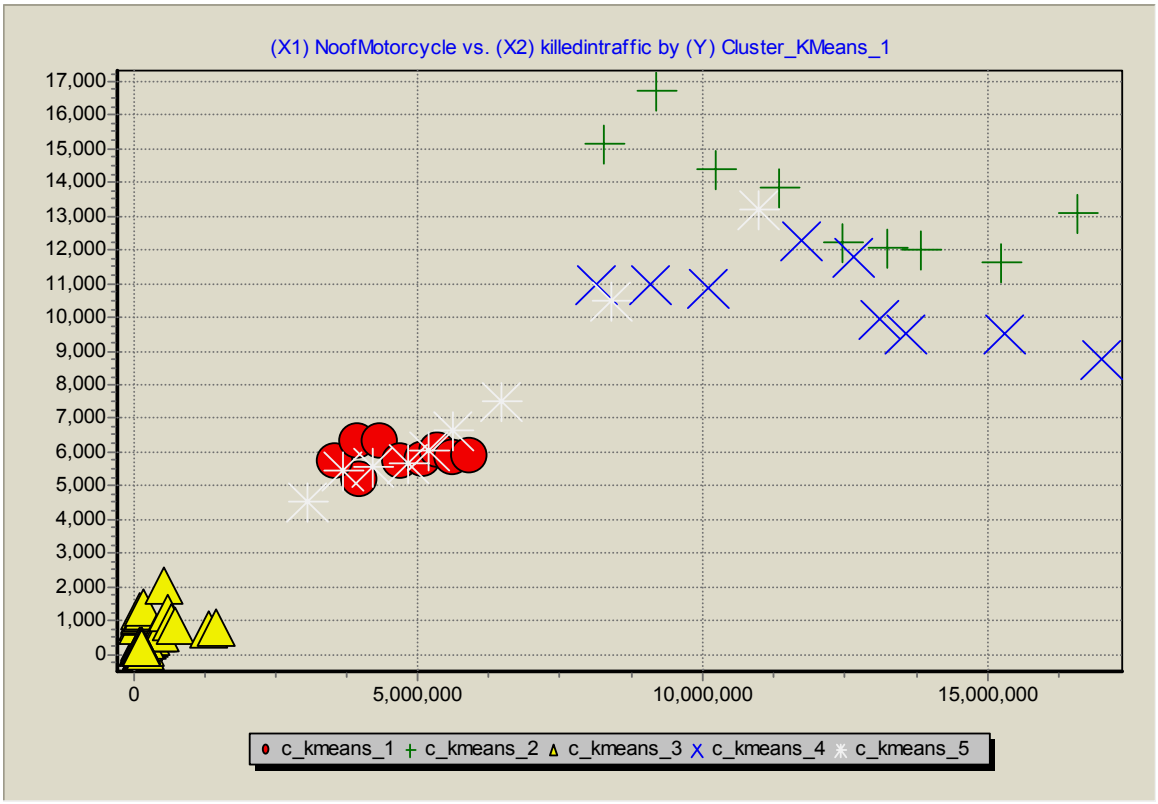

Figure 12. Scatterplot representing the relationship between numbers of motorcycles and the number of people killed using K-means clustering.

To investigate other relationships, selection methods can be applied to change the value represented by the axis. In Figure 13, the x-axis represents the total number of vehicles and the y-axis represents the population. The figure indicates that there is an inverse relationship between the population and the number of vehicles which is peculiar to the ASEAN region. The implication from this is that the most highly populated countries have the smallest number of vehicles (Indonesia and Thailand), while Malaysia (red circle) has the highest number of vehicles but has a smaller population.

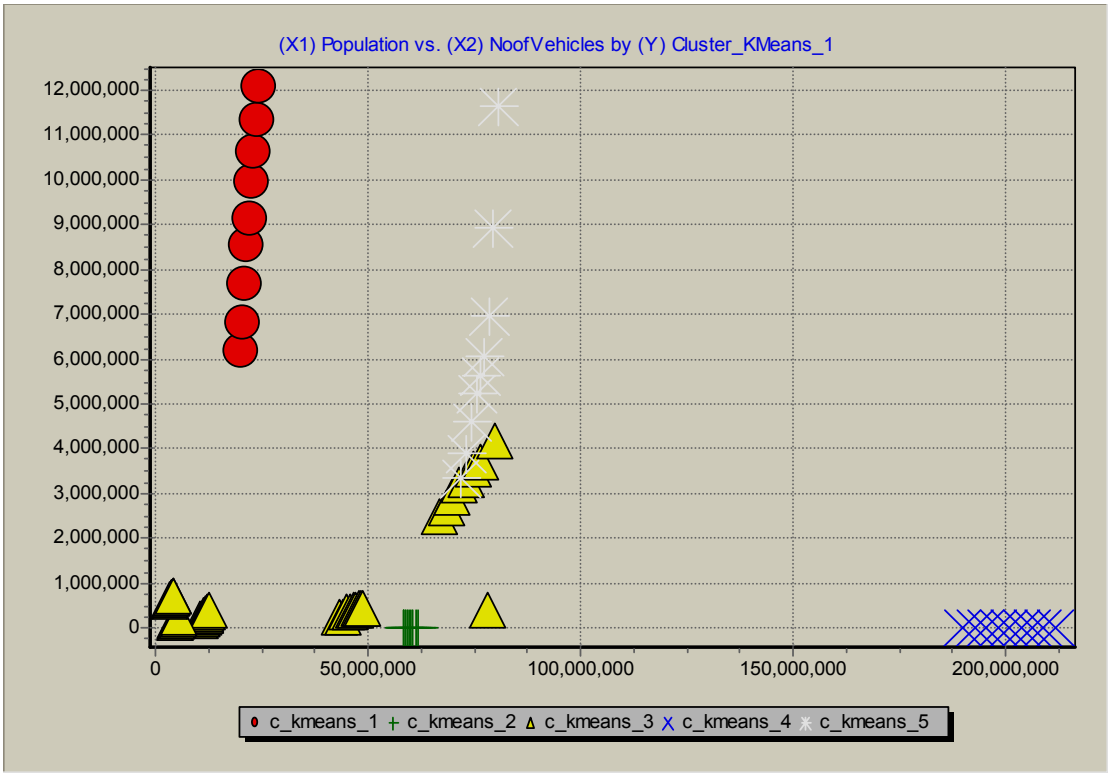

Figure 13. Scatterplot representing the relationship between numbers of vehicles and number of people killed with respect to K-mean clustering.

The same relationship can be explored by using other visualizations, in Figure 14 parallel coordinates were used. The six variables represent the six axes of the PC (population, total number of vehicles, total number of motorcycles, number of people killed in traffic, total number of accidents, and number of injures in accidents). The color scheme of the diagram was defined according to the countries and each color represents one country. The red line in the PC represents Indonesia, which has the highest value in terms of population and number of motorcycles. Malaysia, shown in brown, has the highest number of vehicles and also the highest number of accidents. Thailand, the purple line, has the highest number of people killed (fatalities) and injures due to road accidents. The rest of the lines were not well separated and the countries in these colors require more detailed exploration.

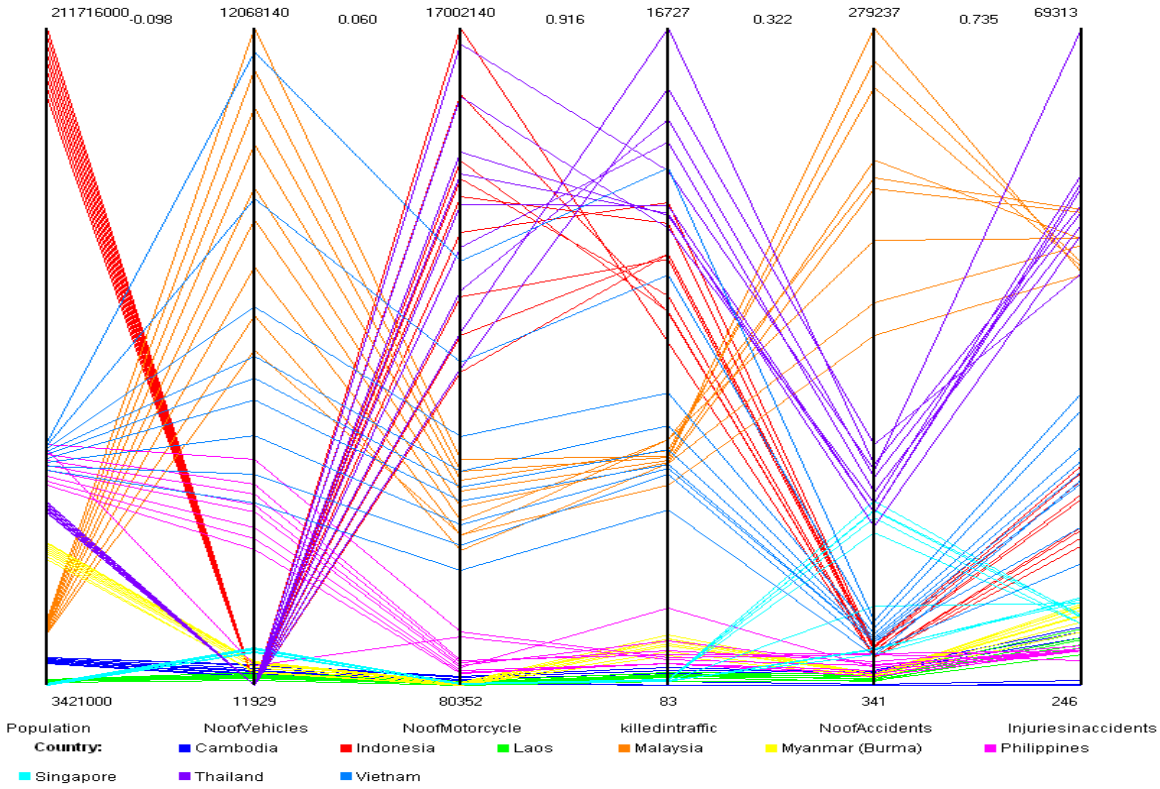

Figure 14. Parallel coordinates showing the relationship between different variables. The variables are represented with axes that are ranged from minimum to maximum.

From the PC in Figure 15, it was possible to infer some characteristics with regards to the situations in specific countries. Singapore (blue line) has the lowest population, number of motorcycles, and people killed in traffic. The rest of the countries are represented by a grey polygon. A comparison of the number of accidents also shows that Singapore has the third highest incidence of road accidents for the countries in the region.

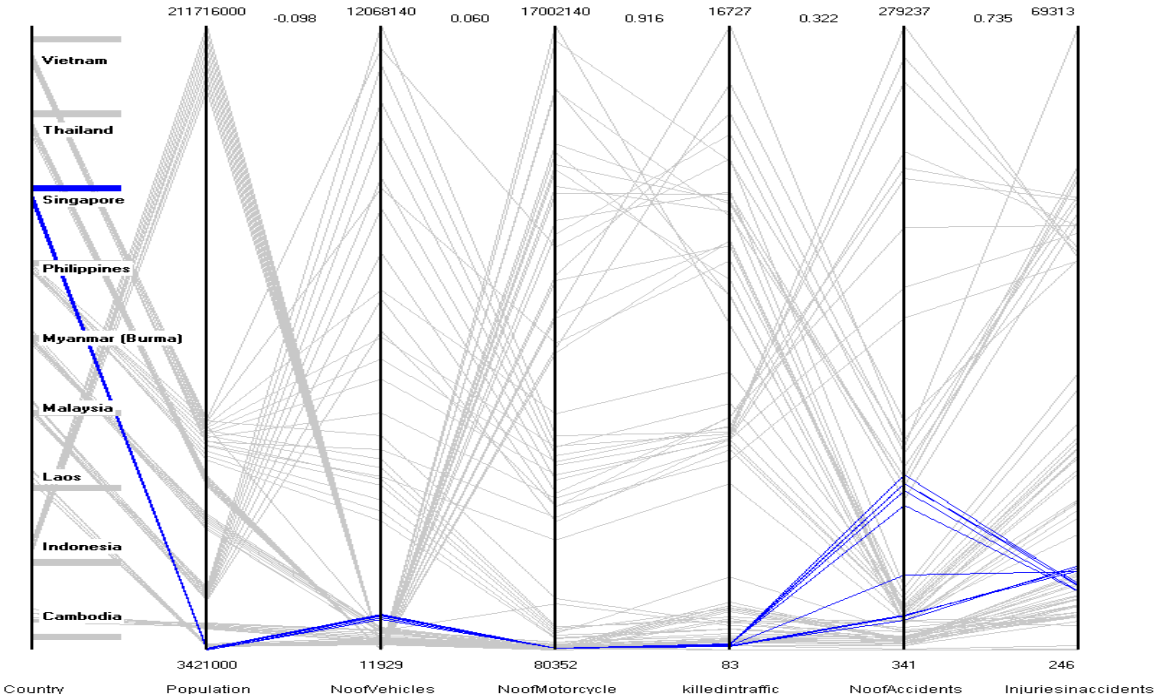

Figure 15. Parallel coordinates showing Singapore as blue lines and the relationship, between different variables. The variables are represented by axes that range from minimum to maximum. The rest of the countries are represented in grey.

6.3. Visual Exploration of Oil Consumption by Road Transport Section

Comparative descriptive statistics were applied in order to characterize groups of countries and their oil consumption from 1990-2003.

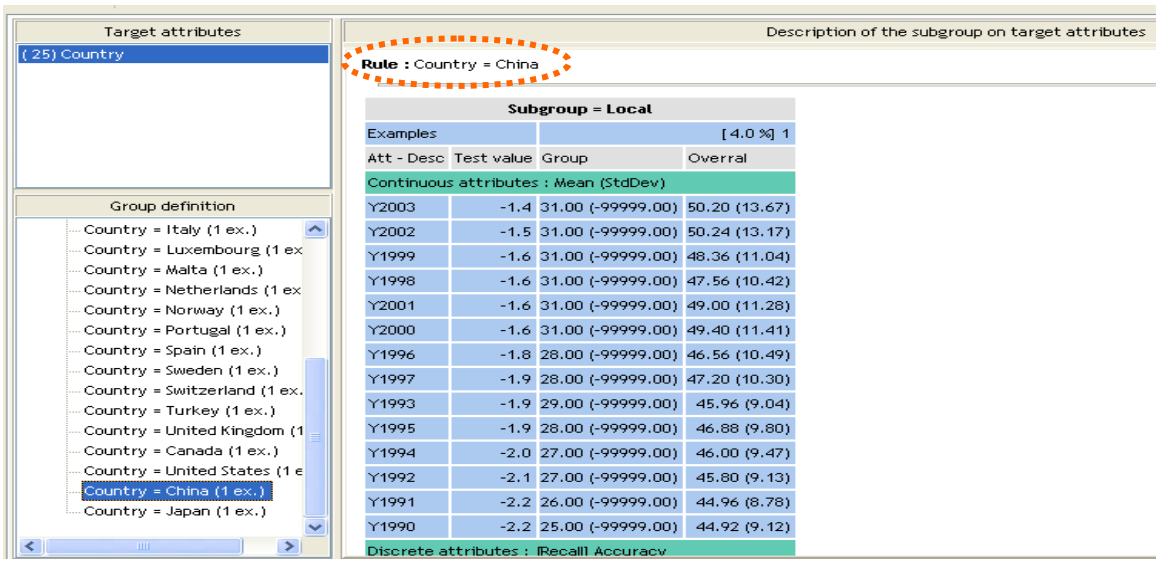

Figure 16. Description of the subgroups on target attributes, *China* was selected from the group definition.

The aim of this exploration task was to show the difference between the countries according to various statistical indicators, such as average (do you means the mean, median or mode when you say average?), standard deviation, and proportion.

The group is defined by a discrete variable (country) attribute. The descriptive statistics are computed on continuous input variables (Y2003, Y1990). In this example, the user can visually explore the data set for the consumption of oil in different countries. In Figure 16, China has been selected in order to gain a better understanding of the consumption of oil in this particular country, using the descriptive statistics.

HAC clustering, in Table 3, was produced from the data set for oil consumption and is explored in Figure 17.

Cluster	Examples in the cluster	Average
C_HAC_1	12	53.1667
C_HAC_2	5	38.4000
C_HAC_3	5	40.2000
C_HAC_4	1	82.0000
C_HAC_5	2	61.0000
All clusters	25	49.4000

Table 3. HAC clustering of oil consumption data set.

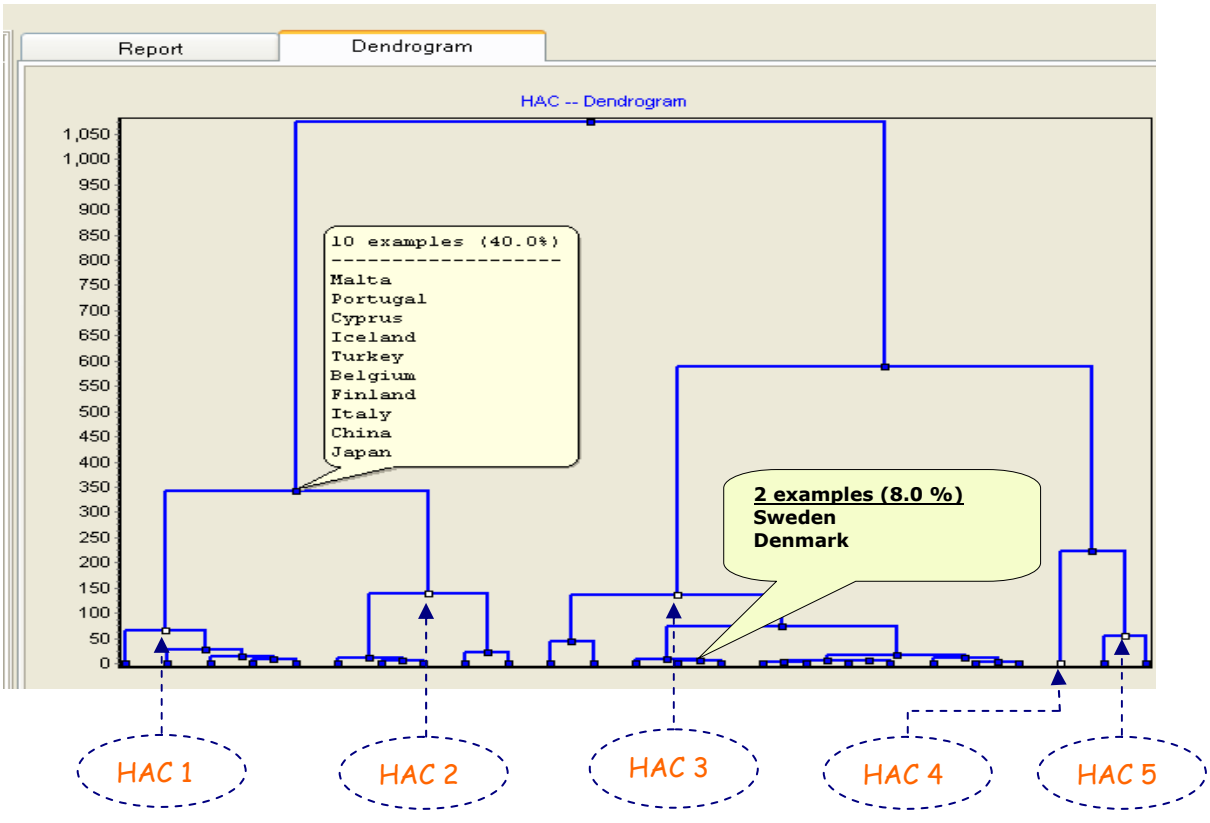

Figure 17. Dendrogram showing 25 countries divided by HAC to five clusters (HAC1… HAC5).

The dendrogram in Figure 17 visualizes the clustering results from Table 3. At the top level of the diagram, all 25 countries are shown as a singleton clusters. According to

the HAC clustering, five clusters were formed. Figure 17 shows that 10 countries share the same cluster. This means that the countries in this cluster are similar in their oil consumption within the road transport sector. This similarity between the countries is shown by the vertical axis which displays a similarity scale. When the similarity of two groups of objects (Malta, Portugal, Cyprus, Iceland, and Turkey) and (Belgium, Finland, Italy, China, and Japan) is 1050 they are merged together to form a single cluster.

6.4. Relationship Between Total Number of Kilometers, Consumption of Diesel and Gasoline

This relationship helps to explain the human behavior towards transportation, and the use of public transport. In this task, linking methods were used, based on Radviz from an energy consumption data set from the OECD countries, containing 18 countries and 612 records for the time period 1965-1998 represented by points in the plot (Figure 18). In Radviz the data features (total number of kilometers, consumption of diesel, consumption of gasoline) are represented as equidistant anchors on the circle.

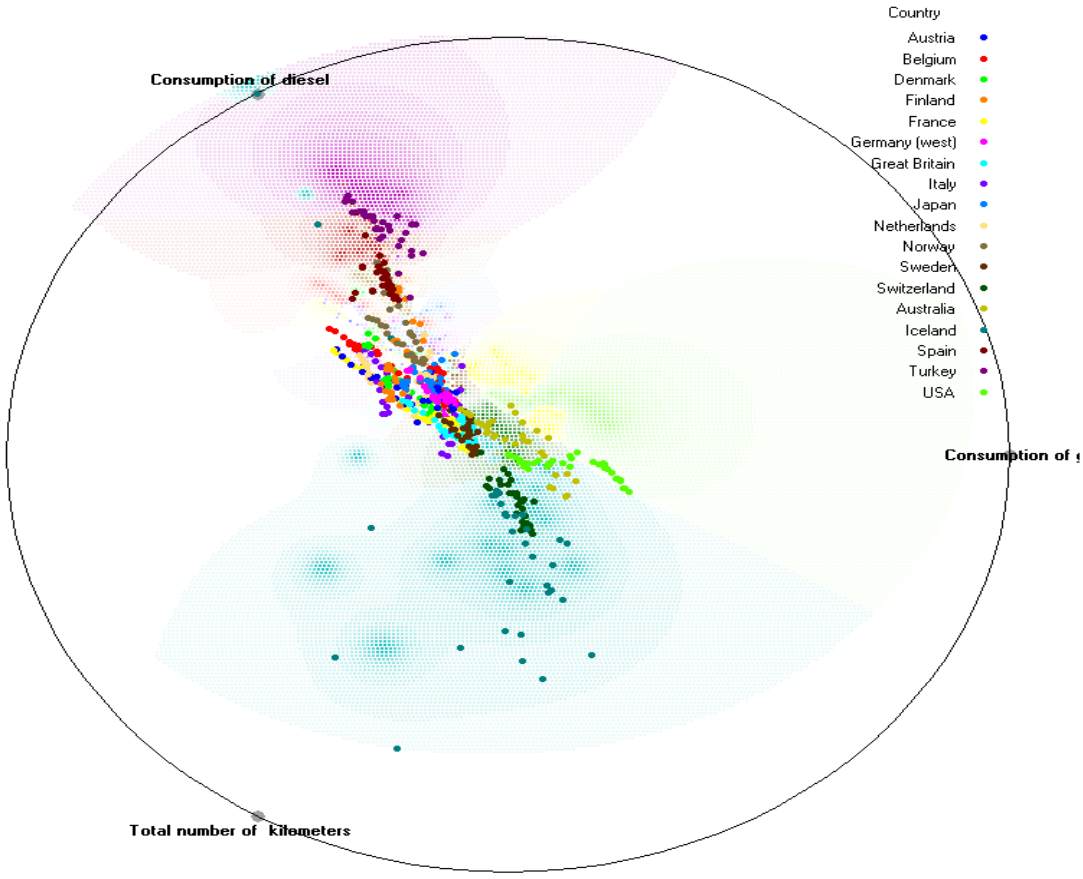

Figure 18. Radviz illustrating three variables, each country is represented by a point.

From Figure 18 it can be seen that:

- All of the countries which have approximately the same or equal values for all three attributes lie close to the center of the circle.

- In the case in which one attribute value is larger than the values of the other attributes, the point will lie close to the point on the circumference of the circle that corresponds to this attribute (green, purple, and teal points).

Countries such as Japan, Australia, United States, United Kingdom, France and Germany, are characterized by a high consumption of diesel and gasoline for transport use and are represented in the centre of the circle. Other countries such as Iceland, Luxemburg, Denmark, and Sweden have lower values in terms of total kilometers and consumption of diesel and gasoline. Countries with high values are to be found lying mainly in the overlapping region within the centre of the circle. By using brushing and linking of the data values in the overlapping region it is possible to link those values to another type of visualization. In Figure 19, the grey rectangle is used to brush and link points in the overlapping region and the selected points are linked to the statistic attributes visualization in Figure 20.

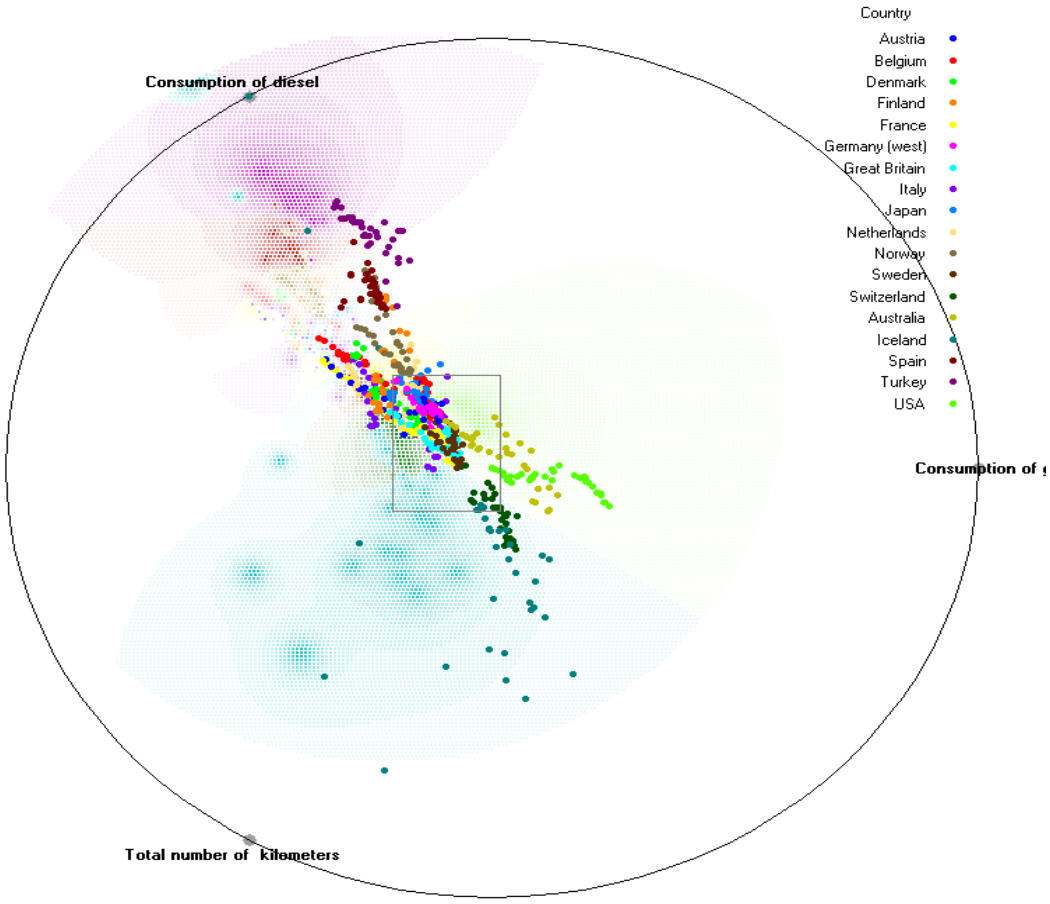

Figure 19. Radviz illustrating brushing of the overlapping region with a grey rectangle.

The statistic attributes analysis is shown in Figure 20 and this visualizes countries as categories plotted in the overlapping area and their occurrences in the dataset. As an example, Austria has a total value of 24 points in the overlapping region, Belgium 17, and so forth, as in the green bar of Figure 20. The statistic attributes provide the minimum, mean, and maximum values of the total number of kilometers.

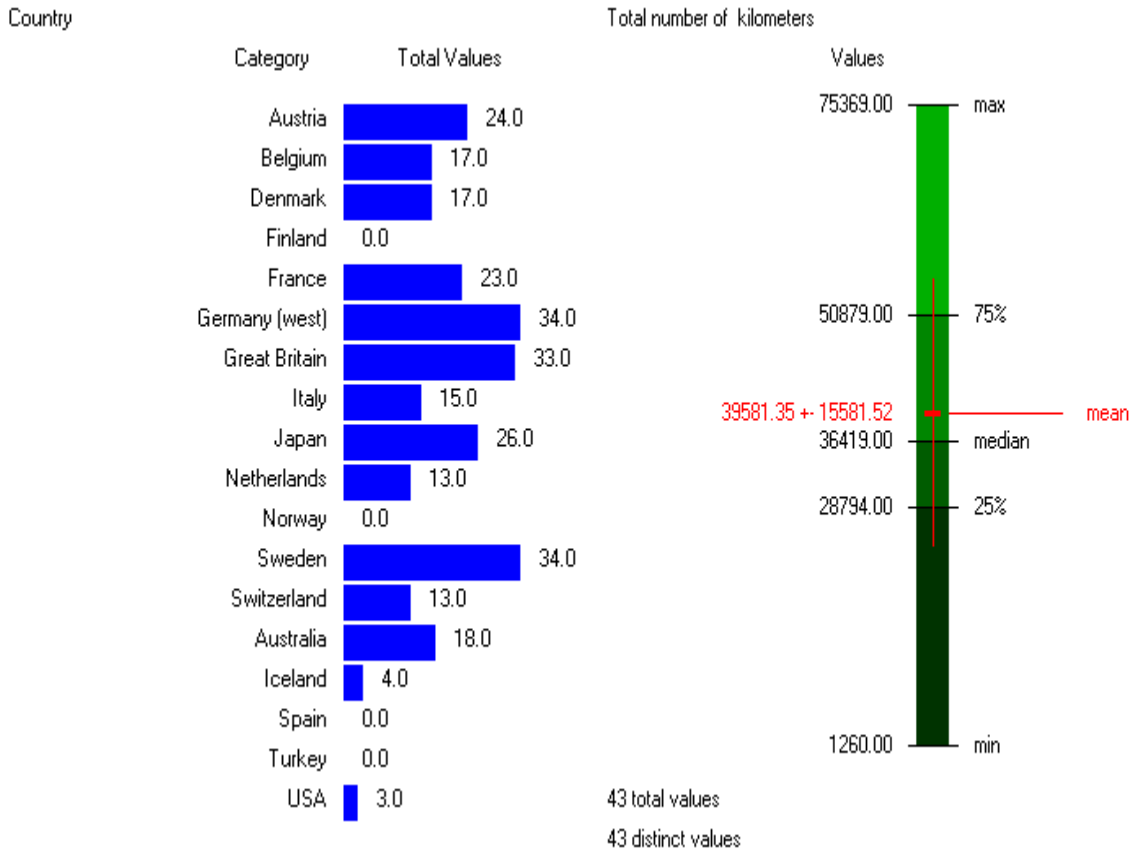

Figure 20. Statistic attributes for overlapping region.

6.5. Exploration of Hospitalized Casualties by Road Use Type

Self-Organized-Map algorithms were applied as methods for data exploration and from which distance matrices were produced. Figure 21 shows the data set for this task from the IRTAD database for a group of OECD countries. The file contains 1919 records and four attributes (country, years 1970-2005, injury type, and traffic participation).The user is able to set the number of the dimensions at the beginning using a SOM classifier interface, in this case the *x*-dimension is 10, and the *y*-dimension is 5. The size of the circle within each cell represents the number of data records in the SOM cell, the larger the circle size the greater the number of records in that cell.

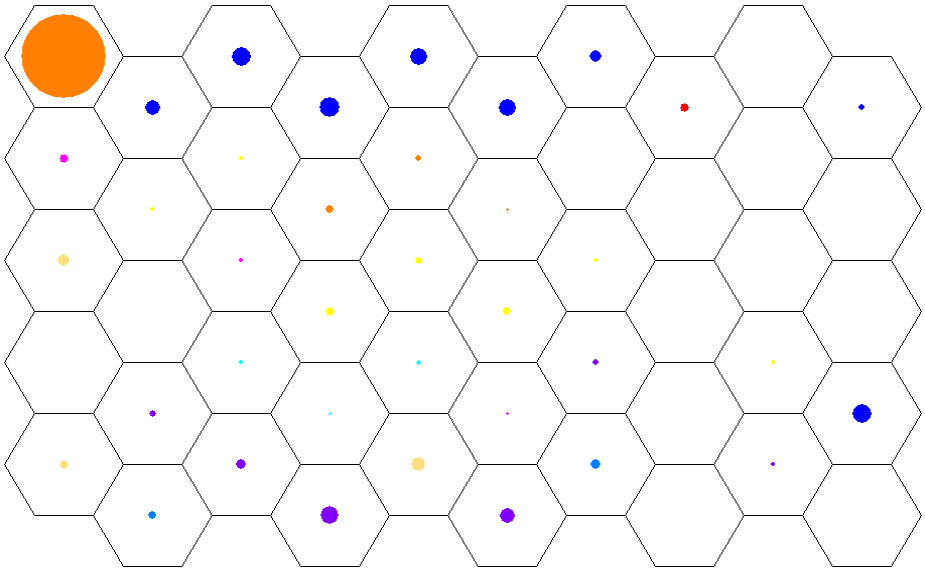

Figure 21. 10x5 SOM matrix with hexagons.

To examine the contents of each cell, the SOM matrix revealed 10 clusters represented by colored circles. This SOM matrix was linked to other visualization methods and the results of this are shown in Figure 22.

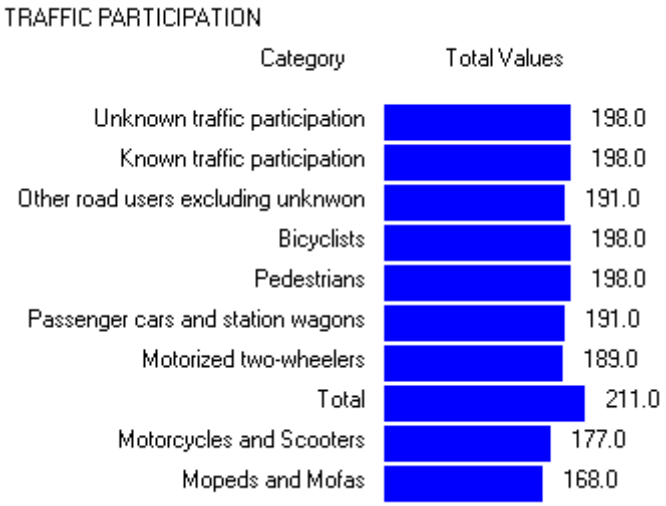

Figure 22. Traffic participants in road accidents linked from SOM matrix.

Figure 22 shows the types of traffic participation in the dataset. In order to be able to see the similarity and distribution of the countries in the SOM (Figure 21), a new distance matrix was created by selecting a subset of the cell using brushing techniques (cells with blue lines in the borders) and the selection is shown in Figure 23. The selection can also be inverted to see the distribution of the rest of the cells. These cells have been selected because the size of the circle in the upper left corner is larger, indicating that there are more data objects in the cell. The distribution of the traffic participation by country is shown in Figure 24.

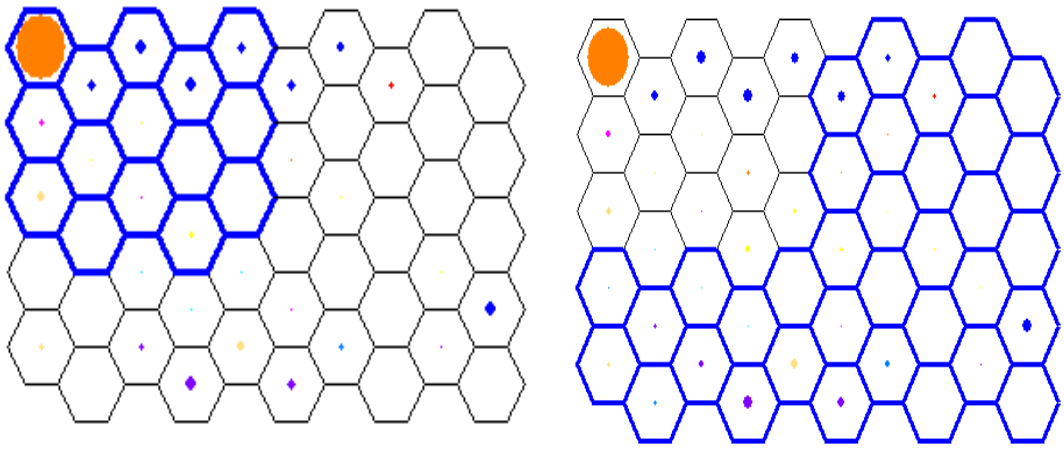

Figure 23. A 10x5 SOM with selected blue cells from Figure 21, the right SOM shows the inverse selection.

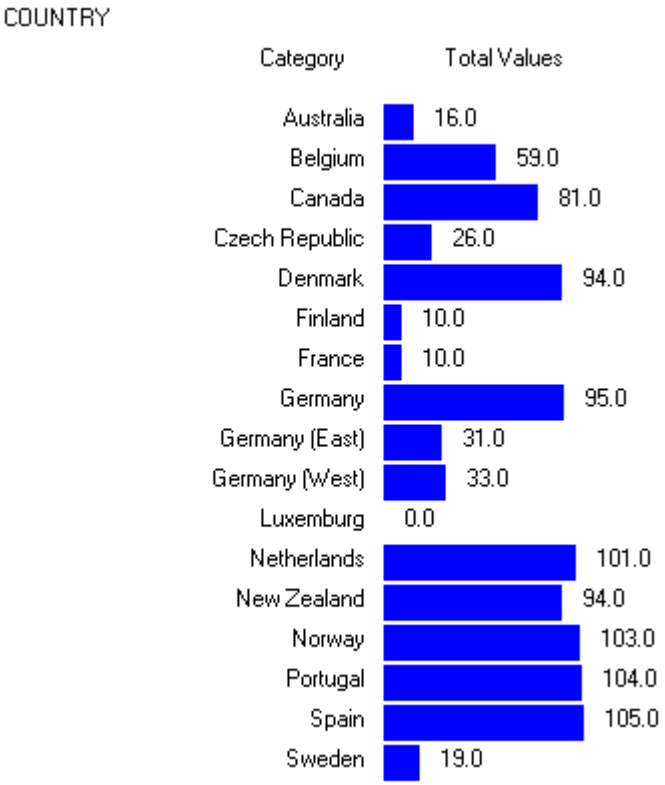

Figure 24. Statistic attributes for selected cells showing the country records within the selection.

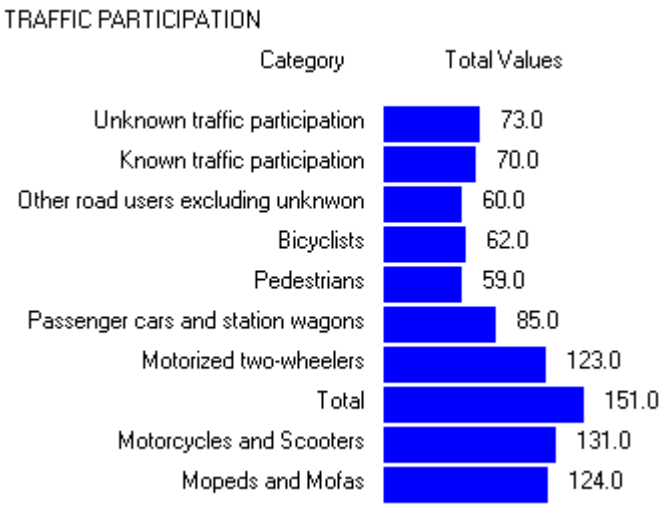

Figure 25. Statistic attributes for the inverse SOM matrix in Figure 23 showing total values for traffic participation.

7. Conclusion

The approach used in this paper integrates visual and automatic data mining. It proved to be of great interest and also very valuable for the purpose of discovering knowledge and relationships in road accident databases. The results obtained assist in the understanding of safety related situations in a region or specific country. This was performed by detecting similar clusters and patterns in the dataset. The road safety expert might be able to estimate and observe relationships that could be hidden. The methods used prove the capabilities of the knowledge discovery tool. However, the low dimensionality of the current dataset limits the exploration task. The VDM methods used proved to be useful in discovering different relationships in the dataset.

The main goal was to find countries which had a high incidence of accidents, people killed in accidents, injures in accidents, highly motorized countries and countries consuming high quantities of diesel and gasoline. Detection of countries with high values for these variables could assist in the identification of particular reasons for the occurrence of such trends (e.g. large population, weak safety regulation, poor road design, or others). Another advantage of this approach is that hypotheses can easily be formulated for future trends. The visualization of different attributes enabled it to be possible to describe similar characteristics for a single country as well as a group of countries in a region and which were visualized within the same cluster. With reference to the main objective, the discovered knowledge was expressed in an understandable way using InfoVis techniques. Future work will focus on geospatial mining of accident locations.

Acknowledgment

The author is grateful to the following people: Urška Demšar, National Centre for Geocomputation, Ireland, for guidelines at the beginning of this research, Susanne Reichwein from IRTAD, and Susanne Gustafsson from Vti for the arrangement of data sets that were used in this article. Special thanks to the referees and editorial board of IJPIS for their helpful comments on improving this article.

References

Abugessaisa, I., Sivertun, Å., and Le Duc, M. (2007). "GLOBESAFE: A platform for information sharing among road safety organizations", in *Proceeding of 9th International Conference on Social Implications of Computers in Developing Countries, São Paulo, Brazil*. CD-ROM ISSN 1981-3945.

Breunig, M., Kriegel, H., Krüger, P., and Sander, J. (2001). "Data bubbles: Quality preserving performance boosting for hierarchical clustering", in *Proceedings of the 2001 ACM SIGMOD Int.' Conf. on Management of data*, pp. 79–90.

Demsar, J., Zupan, B., Leban, G. (2004). *Orange: from Experimental Machine Learning to Interactive Data Mining,* White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana.

Demšar, U. (2006). *Data mining of geospatial data: combining visual and automatic methods*, Doctoral thesis in Geoinformatics, Stockholm KTH Press.

Fayyad, U., Grinstein, G., and Wierse, A. (2002). *Information visualization in data mining and knowledge discovery*, San Francisco, Calif. Morgan Kaufmann; London: Harcourt, cop.

Grinstein, G., Trutschl, M., and Cvek, U. (2001). "High-dimensional visualizations"*,* in *Proceedings of the Visual Data Mining Workshop, KDD'2001.*

Gurubhagavatula, I., Nkwuo, J. E., Maislin, G., and Pack, A. I. (2008). "Estimated cost of crashes in commercial drivers supports screening and treatment of obstructive sleep apnea", in *InternationalJournal of Accident Analysis & Prevention*, vol. 40, no. 1, pp. 104-115.

Hand, D., Heikki, M., and Padhraic, S. (2001). *Principles of data mining*, Cambridge: The MIT Press, cop.

Harris, R. L. (1999). *Information Graphics: A comprehensive illustrated reference*. New York Oxford Press.

Heinrich, J. and Mikulik, J. (2005). "IRTAD–Reliable past and challenging future", in *Proceeding of Vti Road safety on Four Continents*, Warsaw, Poland: Vti, Linköping Sweden.

Hoffman, P. E., Grinstein, G., Marx, K., Grosse, I., and Stanley, E. (1997). "A visual and analytic data mining", in *IEEE Visualization*, vol. 1, pp. 437–441.

Inselberg, A. (1981). *N-dimensional graphics, part I--lines and hyperplanes*, Technical Report G320-2711, IBM Los Angeles Scientific Center, IBM Scientific Center, 9045 Lincoln Boulevard, Los Angeles (CA).

Ioannidis, Y. E. and Christodoulakis, S. (1993). "Optimal histograms for limiting worst-case error propagation in the size of join results", in *ACM Tranactions on. Database Systems.*, vol. 18, no. 4, pp. 709-748.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999). "Data clustering: a review", in *ACM Computer. Survey* vol. 31, no. 3, pp. 264–323.

Keim, D. (2002). "Information visualization and visual data mining", in *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, pp. 1-8.

Nichols, M. J. and Newsome, W. T. (1999). "The Neurobiology of Cognition", in *Nature*, no. 403, pp. C35-C38.

Pang-Ning, T., Steinbach, M., and Kumar, V. (2006). *Introduction to data mining*, Boston: Pearson Addison Wesley.

Poole, D., Mackworth, A., and Goebel, R. (1998). *Computational Intelligence: A Logical Approach*, New York Oxford University Press.

Rakotomalala, R. (2005). "TANAGRA: A free software for research and academic purposes", in *Proceedings of EGC'2005, RNTI-E-3*, vol. 2, pp. 697-702.

Voigt, R. (2002). *An Extended Scatterplot Matrix and Case Studies in Information Visualization*, Master's thesis, Hochschule Magdeburg-Stendal. Germany.

Shneiderman, B. (1996). “The eye have it: A task by data type taxonomy for information visualizations”, in *Proceedings of IEEE Visual Languages* , pp. 336-343.

Soukup, T., Davidson, I. (2002). *Visual data mining techniques and tools for data visualization and mining*, Wiley, New York.

Stuart, K., Mackinlay, D., Shneiderman, B. (1999). *Readings in information visualization: Using vision to think*, San Francisco, Kaufmann, cop.

Tufte, E. (1986). *The visual display of quantitative information*, CT, USA Graphics Press, Cheshire.

Gitelman, V. and Hakkert, A. S. (1997) “The evaluation of road-rail crossing safety with limited accident statistics” in *International journal of Accident Analysis & Prevention,* Vol. 29/2, pp 171-179.

Yannis, G., Golias, J., and Kanellaidis, G. (1998). “A comparative analysis of the potential of international road accident data files”, in *International Association of Traffic and Safety Sciences Research*, vol. 22, no.2.