TEMPORAL VARIATION OF ROAD ACCIDENT DATA CAUSED BY ROAD INFRASTRUCTURE

Amirhossein Ehsaei
Center for Infrastructure Research, Department of Civil Engineering, University of Louisville, Louisville KY, USA, e-mail: a.ehsaei@louisville.edu

Dr. Harry Evdorides
School of Civil Engineering, University of Birmingham, Birmingham, United Kingdom, email: h.evdorides@bham.ac.uk

Submitted to the 3rd International Conference of Road Safety and Simulation, September 14-16, Indianapolis, USA

ABSTRACT

Road accident data are collected annually in the UK using the STATS 19 accident report forms. The data is then stored on databases for subsequent analyses. The purpose of this paper is to present a methodology which may be used to extract knowledge from such databases in an automated manner using data mining techniques. The analysis examined accident causation due to road infrastructure features over a period of 5 years. The data mining system used was WEKA. Accident frequency histograms for each year were computed together withthe probabilities of occurrence of a fatal/serious injury accident for different circumstances. . Using Bayesian classifiers and an appropriate search algorithm in WEKA it was possible to analyze the data considered by producing rules that describe the interrelationship between different data. The success of this methodology was demonstrated by the agreement between the results and the matching road policies adopted in this period by the road authorities concerned. For example, results show that generally it is more probable that an accident will occur in daylight than in darkness. Another finding of the study is that the possibility of fatal accidents in darkness, when lights are unavailable, is twice as high as the possibility of serious injury accidents. It was also found that the probability of fatal and serious accidents occurring on wet surfaces had declined over the five-year period under consideration.

Keyword: accident causation, STATS 19, data mining, WEKA.

INTRODUCTION

Approximately 3000 road users every year are killed on the road networks of the UK and 32,155 people were killed or seriously injured in road traffic accidents (RTAs) in 2005. However, this number is 6% lower than the year before and about 33% less than the average for the period years 1994 to 1998according to the Department of Transport (2001-2006). RTAs waste 4% of the Gross National Product (GNP) of countries around the world (WHO, 2010). **(Please look at your referencing system and how you cite your references)** RTAs have three main causal

factors: vehicles, pedestrians and road infrastructure: PIARC (2007). The occurrence of accidents caused by infrastructure varies between different counties in the UK and different types of road environment. Since accidents are systematically recorded by the police using an appropriate system (STATS 19) , it is possible to trace the causes of accidents, including road infrastructure. But with this massive volume of data, it is not easy to examine the relationship between accidents and their contributory factors.

In order to be able to achieve a good understanding of the massive databases containing data from annual accidents, understand their details and discover the hidden trends in the information, data mining techniques have proved to be useful. Data mining is searching through data to find connections which were previously unknown: Exforsys (2010). The intention of this paper is to show the development of a methodology which uses artificial intelligence and data mining techniques to help road engineers understand the effect of road infrastructure on the occurrence of accidents over a specified period (e.g. 5 years).

Road infrastructure includes all features which are required for motorized traffic, including traffic safety areas and noise protection elements.: Link et al. (1999). However the data required for investigating the role of road infrastructure in accident causations is not always available. Because of this, only the aspects which are mentioned in the database are considered.

LITERATURE REVIEW

In order to reduce the number or severity of road accidents caused by road infrastructure, appropriate measures need to be implemented. It is however important to identify or even predict the location where accidents may occur. . The processing of large accidents databases may assist in this: Wu and Heydecker (1993). Identification of the locations where accidents happen was a key issuein the study. Althoigh the site of an accident can be identified by appropriate codes ,there were no explicit relationships between the accidents which happen in the same area i.e. the area with same code. Consequently before undertaking the analysis, locations should be identified so that cases of several accidents occurring in similar circumstances can be easily recognized: Heydecker and Wu (1999).

Data mining has helped to reveal patterns in road accident databases. These databases contain many rows and columns of data and, by means of modern tools for collecting and storing data, it is now possible to analyze these data and put the results to good use: Exforsys (2010). With the help of concepts such as data mining, road safety has improved dramatically and data mining has been used in many traffic related studies: Beshah and Hill (2008). Data mining has been used to investigate the importance of road related factors in the severity of accidents, on the basis of road accident data from Ethiopia using predictive models: Beshah and Hill (2008). In that study, in order to predict the severity of accidents, they used classification models which were based on decision trees that revealed driver behavior and the connection between roadway and weather conditions and accident severity.

A study on Finnish roads applied large scale data mining methods to overcome the difficulties of analyzing a multidimensional and heterogeneous database with data for 83,000 accidents between 2004 and 2008. The study data were collected by the Finnish Road administration:

Ayramo et al. (2009). The outcome of this study shows that the use of descriptive data mining techniques makes it easier to extract knowledge form a database.

DATA SELECTION

The data for road traffic accidents used in this work was taken from the UK data archive (UKDA) (http://www.data-archive.ac.uk/findingData/radTitles.asp). It focused on the road traffic accidents which occurred between 2001 and 2006. The data come from police reports recorded on appropriate forms (STATS 19).. The method of collecting data in Great Britain using STATS 19 is as follows. Details of every accident on public roads are recorded on the forms by police officers. Within each county of England, these data are collected by the Local Processing Authority (LPA), which is the central unit for this county. The data should be validated first and the LPA provide the data to the DfT and also to local highway authorities: DETR (2001). Data are annually compiled and submitted to UKDA and DfT on electronic media. For the public and researchers, data can be accessed through the UKDA website.

For each year, three different categories of data are available. Among these categories only the items which are related to road infrastructure were considered:

- Accident: Includes time/day/month/year of accident, accident severity, class/ type/ number of road, light conditions, road conditions, junction detail
- Casualty: Class/ type/ severity of casualty
- Vehicle: Hit object in/off carriageway, skidding/ overturning, position of vehicle

From these, variables related to Junction Details were considered in this study. However the methodology can be used for any variable.

According to the data provided by UKDA for each year of study, accident data are available for 41 counties in England, 4 counties in Wales and 8 counties in Scotland. In order to adjust the study to technical limits, e.g. computer capabilities, a county is selected on which the study concentrates.

.

DATA ANALYSIS

, The database structure should be known, in order to describe the correlation between the datasets. After data mining, these variables should graphically show the temporal variation of each defect in a five-year period of study..

Severity of Accidents

Some of the data are negligible and these must be distinguished from the rest at this stage. For example, regarding the severity of accidents, all three types (fatal, serious and slight) may be included or slight injuries may be excluded. As the social cost of fatal and serious injuries accidents is much higher than that of slight injuries, only fatal and serious injuries were considered for this study: Mohan (2002).

Junction Detail

The data which are related to road geometry within road accident databases are concerned with the junction detail, as the accident rates are different between different locations of the road, such as roundabouts and multiple junctions.

DATA SET USED

Data Set Selection

Among the databases retrieved from the UKDA for the accidents between 2001 and 2006, there are three sets of data available concerning the distinctive variables of Accident, Vehicles and Casualties. The accident database has 30 variables, the vehicle data base has 23 and the casualties' database has 16. Some of these variables are the same in all databases, such as accident year, but they also have differences.

The accident database holds details such as:

- Year, month, day and time of the accident
- Junction details
- Accident severity and the number of cars involved in the accident

The vehicle database includes details of the vehicles involved, such as:

- Vehicle maneuvers
- Direction of movement of vehicles at the time of impact (to/from)
- Skidding/Overturning

The databases of casualties give details of the followings:

- Age/sex/class of casualty
- Severity/Type of casualty
- Pedestrian movement/location/direction (if any)

The specific purpose of the present study is to reveal the temporal variation among accidents which are caused by road infrastructure. The only database which holds information relevant to the objective of this study is the accident database and so this alone will be used.

Selection of study area

One typical county was chosen. It was essential that this choice should show the diversity of data across the UK and in the end Greater Manchester was thought suitable. This county has the second highest population of any across the UK, excluding Greater London (see Table 1).

The county of choice for this study has the following population and density. Greater Manchester, as the third largest county of the UK, had over 13,540 killed or injured in road accidents in 2004.

This cost the community of Greater Manchester about £825m. Even though this county has reduced the number of KSI, in order to meet the targets of 2010, it must continue a coordinated approach to solving the problem of road safety: GMLTP (2005).

Table 1: Population of the selected county comparing to Greater London

Rank	County	Type	Population	Density
1	Greater London	Administrative area	7,611,900	4,851
3	Greater Manchester	Metropolitan county	2,573,500	2,017

Source: Office for National Statistics

For this county data on the killed and serious injuries (KSI) and slight injuries as well as the KSI for children, were considered and shown in Table 2. The KSI among young people and children in particular raises the cost to society. The government had a plan to reduce the number of road accident casualties by the end of 2010; although Greater Manchester is close to the proposed target, further investigation is being made to improve road safety through other measures, such as road infrastructure.

Table 2: Reported casualties in Greater Manchester for 2008

Greater	Population	Child KSI	All KSI	Slight	Total
Manchester	2573498	139	843	9038	9881

Data Set Used for Analysis

Database for the Greater Manchester county and for each year were formed. The data analysed concerned the severity of the accidents in connection with the location of accidents defined as follows:.

- Not at junction or within 20 meters of one
- Roundabout
- Mini-roundabout
- T, Y or staggered junction
- Slip road
- Crossroads
- Multiple junction
- Private drive or entrance
- Other junction

SOFTWARE

There are various software programs available for the purpose of data mining. The choice for this study, however, was constrained by availability and simplicity of use. In the choice of software, the most important factor was the free use of the software, thus exempting commercial software packages. The software programs which are mostly used by scientific researchers are the R project: Ihaka and Gentleman (2010), WEKA: University of Waikato (2010), KNIME: Bioinformatics and Information Mining (2010), Rapid Miner: The Data Mine (2010). All these software packages are open-source and freely available to use and develop for special purposes. The software of choice should include the relevant techniques required for this study. For the present study, WEKA was proposed.

WEKA (Waikato Environment for Knowledge Analysis) is an open source and popular form of data mining software written in Java environment: University of Waikato (2010). WEKA was developed in the University of Waikato in New Zealand. It is the software proposed for undertaking the process of data mining of the UK accident data. It is essential for the software to have the capacity to perform the required data analysis and predictive modeling, such as classification. It is also important that the result of the analysis can be visualized and shown with a graphical interface.

DATA EXPLORATION

Visual exploration of the data is the first step of data mining and aims to provide an insight into the database, using information visualization techniques. For effective data mining, the human factor has an important role. It is essential to join flexibility and creativity to the great ability of modern computers to process data. Visual data exploration aims at integrating the users in the data mining process by applying their understandings of data sets. Data analysts should gain an insight before drawing conclusions from the data set: Keim (2001).

The approach of data exploration is to understand the data before the analysis process begins. The accident data base contains many variables derived from accident report forms. Visual data exploration before data mining has several advantages: Keim (2001):

- It makes it easier to deal with heterogeneous data
- It creates an insight into the data set
- Unlike the actual data mining, it requires no understanding of complex mathematical concept or statistical algorithms

Preparing the Data File for Data Mining

Data sets contain many variables which can complicate the data mining process and make it time-consuming. The spreadsheets retrieved for the UKDA are huge databases and it is not possible for WEKA to process and analyze them within the given time and with the help of normal computers. Therefore these databases should be transformed into smaller and more efficient databases. New databases should hold only the necessary information.

The process of extracting, transforming and loading (ETL) of data converts the original data into a data set appropriate for the mining process. ETL is the process of converting bulk data into aggregate form. These results in a more efficient data mining process: Huebner (2008). For the purpose of this study, ETL follows the steps listed below:

- Extraction of data from the UKDA
- Transforming data into smaller volumes and separated databases for each county, it may require some other modifications
- Loading transformed databases into WEKA

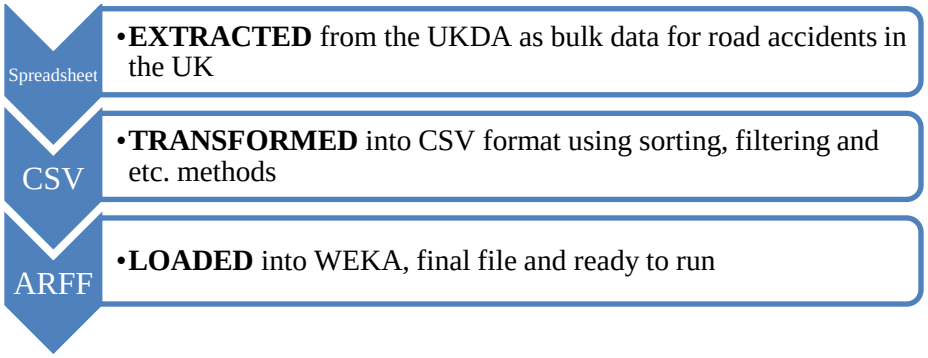

Figure 1: Transformation steps in detail showing different file formats

Figure 2: Transformed file being loaded into WEKA

Visualization

In the pre-processing stage, the software shows a histogram of the frequencies of variables in a box. This is a good tool for validating the data analysis by comparing them to the available diagrams issued by the DfT.

As an example at this stage of the study, the WEKA algorithms were used to extract accident data with regard to a simple feature such as lighting conditions. .
Road lighting reduces injury accidents and : Wanvik (2009).

Table 4 (produced using WEKA)shows that most accidents occur the presence of light, either in day time or darkness.

Table 3: Greater Manchester accident statistics per annum (derived from the accident databases)

Light conditions	Number of accidents per annum – Greater Manchester				
	2001	2002	2003	2004	2005
Day light – lights present	7285	6759	6642	6244	5803
Day light – no lighting	574	534	1112	459	685
Day light – lighting unknown	197	167	444	116	107
Darkness – lights lit	3044	2946	2720	2837	2505

Darkness – lights unlit	33	43	20	34	36
Darkness – no lighting	58	48	60	47	61
Darkness – lighting unknown	65	55	42	30	32

The variation of accidents caused by the light conditions of the road is shown in bar-chart below:

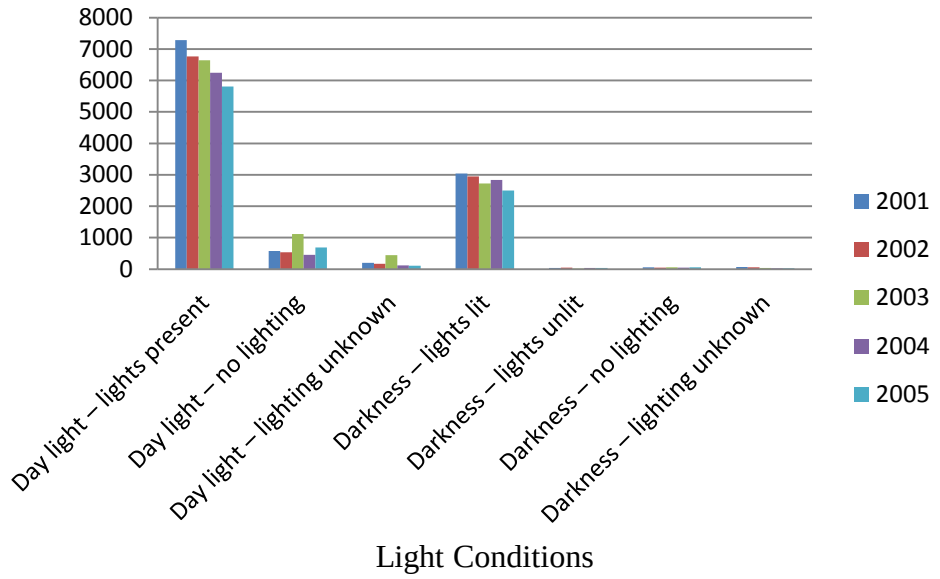

Figure 3: Greater Manchester variation of accidents caused by the light conditions

DATA MINING

The aim of this research is to develop a methodology which facilitates an understanding of the effects of the contribution made by road infrastructure to road accidents over time (temporal). For the purposes of this study, a classification technique has been chosen to generate the desired results.

Classification

Classification is a method in WEKA which can be used to analyze data and recognize patterns. In order to undertake classification, a classifier must be constructed: Friedman et al. (1997). WEKA has such classifiers built in. In order to select the best technique according to the needs of the study, with a series of trials, the Bayesian Classifier was identified as the best algorithm available in WEKA.

Bayesian Classifier

Bayesian networks provide an appropriate language and efficient machinery to represent and manipulate assertions of independence.

Bayesian networks are used to reason under uncertain conditions using probabilities. Bayesian networks consist of the following: Jensen (2009):

- a set of variables and a set of directed edges between variables
- variables which have a limited set of mutually exclusive states
- variables which are derived from an acyclic graph together with directed edges

With the help of Bayesian networks, acyclic graphs are directed and allow an efficient and effective representation of the joint probability of distribution over a set of random variables: Friedman et al. (1997).

A Bayesian network or BayesNet offers four different methods of undertaking the classification process. These are used to train the data to predict their own future trends. The classification technique based on the BayesNet classifier is aimed to generate a model which fits the similarities of attributes and class labels of the input data. To achieve this aim the classifier employs a learning algorithm. If a database consists of 100 instances, it is not logical to base 100% of it on training or testing. By splitting the data into 66% for training and 34% for validating the outcome this problem may be overcome: Reutemann (2010). Validation is an important part of the learning, because training the dataset is not enough and results need to be verified on a portion of the data set. This can be done through the 'classify' tab and in test options. When the percentage split is 66%, WEKA divides the data set into training and test sets.

Data Analysis Using Bayesian Classifier:

Using WEKA, acyclic graphs of the probabilities can be produced. These graphs are the outcome of a classification process using a BayesNet classifier. The figure below shows the correlation between variables to be identified. The order of variables shows the probability of a fatal or severe accident in any of the conditions related to the road infrastructure.

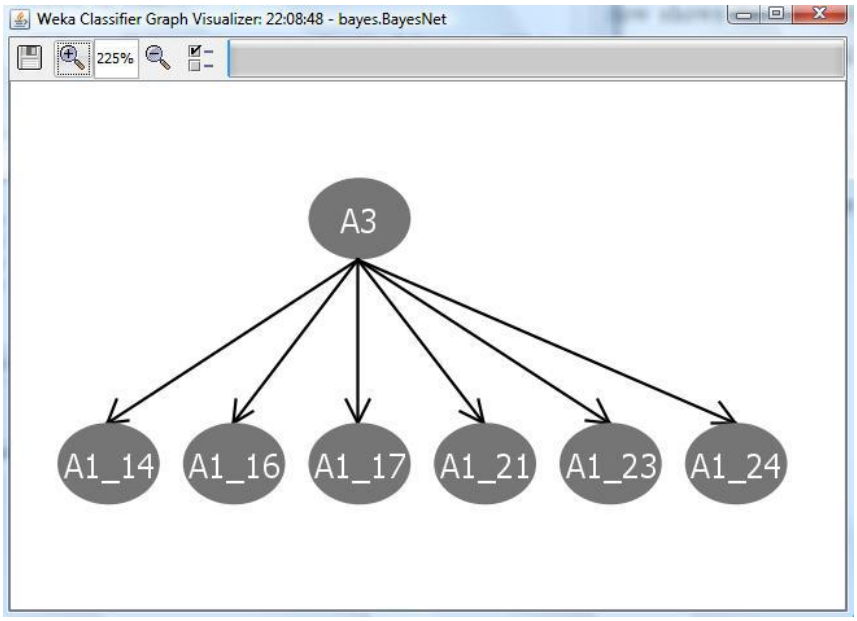

Figure 4: Using a Bayes network for the analysis

In order to simplify the work, unnecessary attributes were removed using a filter in WEKA. Using this filter does not eliminate these attributes, but simply hides them from the analysis since they are not related to the infrastructure of the road. When the graphs are viewed, WEKA also supports a tool to edit the BayesNet graph. The following can be carried out with this option:

- editing the Bayesian network manually
- editing the network structure manually
- editing the conditional probability tables

Can you show an example of the above figure using the data you considered? So instead of showing A3, say for example KSI accidents.

RESULTS

The probabilities of major road infrastructure defects occurring together with accident severity were analyzed. These probabilities were calculated using the Bayesian classifier and were not changed with the BayesNet editor as the ideal probability of occurrence of accidents in different situations was unknown before the study.

It was explained that it was necessary to find the major road infrastructure defects and focus the study on them so that the study could be completed within the given time. The results of the following variables are shown for this study:

- Accident severity
    - Causing fatalities
    - Causing serious injuries
- Junction detail
    - Not at junction or within 20 meters
    - Roundabout
    - Mini-roundabout
    - T, Y or staggered junction
    - Slip road
    - Crossroads
    - Multiple junction
    - Private drive or entrance
    - Other junction
- Special conditions at site
    - None

Efforts were made to limit the scope of the study to road infrastructure only and thus the special conditions at the site of the accident are limited to none, which means there should not be any technical problem with road side features such as traffic lights.

The following table shows the results of classification for the probability of accidents happening in special conditions of a site in year 2001 in Greater Manchester. This table is an example of

very low percentages being shown for the effect of road works, defective road surfaces, etc. on the occurrence of accidents.

Table 4: Probabilities of accidents occurring in special conditions on site

Special Condition	None	Road Works	Defective Road Surface	Obscured Signs
Fata	0.976	0.011	0.004	0.001
Serious Injuries	0.98	0.01	0.005	0.001
Slight Injuries	0.978	0.013	0.003	0.002

In the rest of the database the probability of accidents in the following years were extracted from the result of Bayesian classification. The outputs of the analysis for the county of Greater Manchester can be seen here (where?).

It should be noted that these probabilities are generated to show and compare the possibility of a fatal or serious injury due to an accident in any of the locations. The probability may be shown in another way, which presents a comparison of the likelihood of a fatal or a serious injury in one place, but this is outside the purposes of this study.

When the high-risk infrastructures are identified then extra care can be taken in both design process and improvement plans. With an investigation on the probabilities of fatal accident in relation to junction details throughout the period of analysis the highest rates are constantly at:

- Not at a junction or within 20 meters
- T, Y, or staggered junction
- Cross road

The following Table and graph represents the results for the whole period:. (Please chose either table or graph but not both).

Table 5: Risk of fatal and serious accidents for the 5-year period in Greater Manchester

Junction	**2001**		**2002**		**2003**		**2004**		**2005**	
Detail	Fatal	Serious	Fatal	Serious	Fatal	Serious	Fatal	Serious	Fatal	Serious
Not at junction or within 20 meters	0.6141	0.4677	0.629	0.467	0.637	0.47	0.637	0.488	0.642	0.491
T, Y or staggered junction	0.2258	0.304	0.213	0.3	0.199	0.295	0.207	0.288	0.205	0.287
Cross road	0.0573	0.089	0.065	0.09	0.061	0.078	0.059	0.084	0.052	0.082

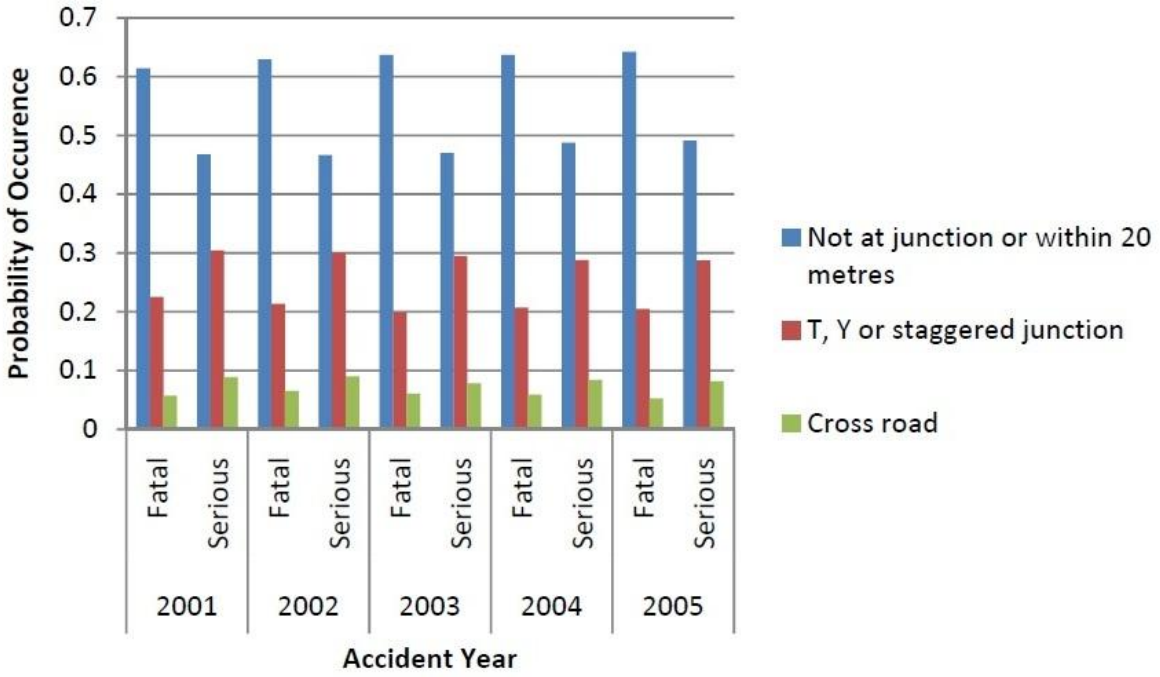

Figure 5: The occurrence of fatal and serious injuries in relation to junction details in Greater Manchester

The following diagrams show the temporal variation of the accidents through the 5 year period of the study:

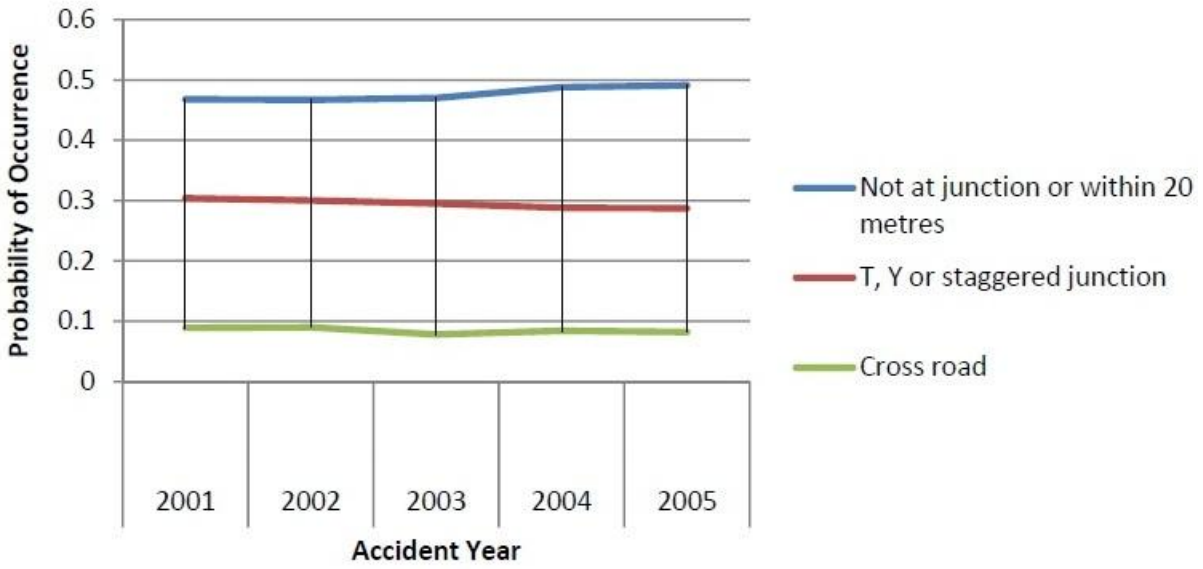

Figure 6 : Temporal variation of fatal road accidents in relation to junction detail in Greater Manchester

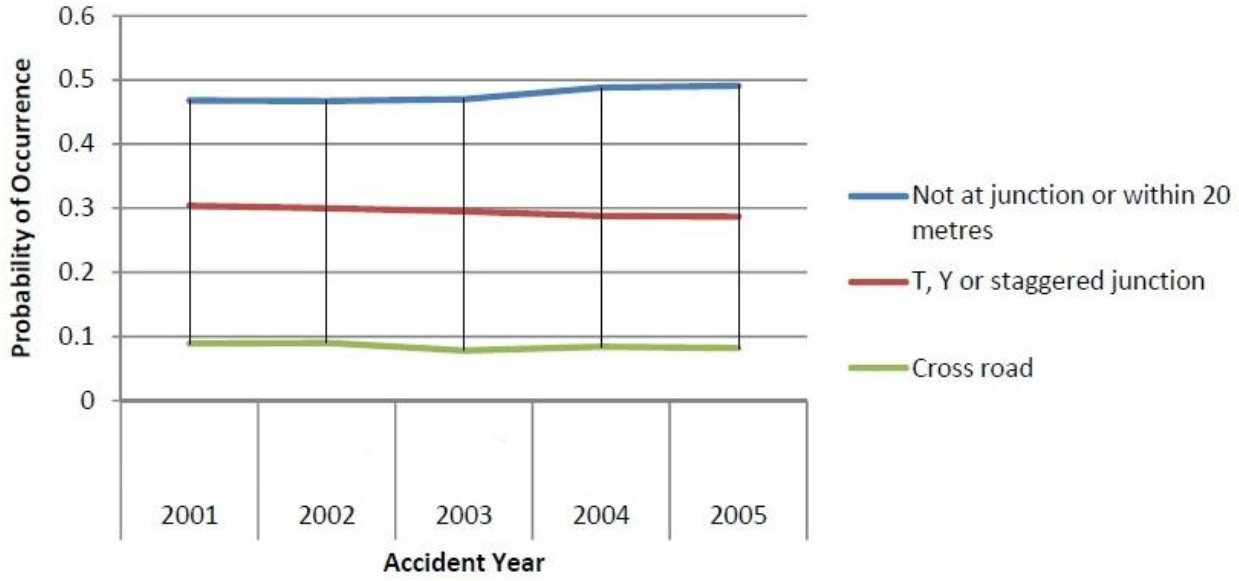

Figure 7: Temporal variation of serious road accidents in relation to junction detail in Greater Manchester

What's the difference between figures 7, 8 and 9?

CONCLUDING DISCUSSION

This paper presented a preliminary investigation on how data mining techniques may be used to extract knowledge from accident databases by using the freeware software WEKA. It has been found that is not necessary to use complex methods and data mining techniques to produce reliable results. In other words, the sophistication of the procedures does not affect the results. With the application of data mining, the hidden relationships and meaningful patterns may be revealed. This pattern can be very valuable for assessing whether road safety measures applied to a road network throughout a period have been effective or not.

In addition, using a data mining techniques the computed probability of a fatal or a serious accident happening in specific conditions may be produced by using a Bayesian method

More specifically, the work presented here showed the probability of an accident at different locations in the road network (such as junctions) can be computed to define a high risk point for fatal and serious accidents can be computed. In a preliminary analysis it was found that three of such locations carry the greatest risk of accidents and so further investigation and analysis was limited to these locations. The highest possibility of an accident is at sites which are "not a junction or within 20 meters". Trends show that the probabilities of such accidents have increased throughout the five-years under reviews. Therefore measures for accident prevention for such locations are needed.

The computed possibility of accidents at T, Y and staggered junctions has slightly declined over the five year period. This is in agreement with the expected results from the implementation of measures varying from warning lights to speed humps were introduced with the aim of reducing the speed of entrance to the junction.

Crossroads, which rank third as possible sites of fatal and serious accidents, hold less than 10% risk. Crossroads have, however, a high risk of accidents involving pedestrians: Pai et al. (2004). The results of a similar study conducted by the Department for Transport indicate that accidents at crossroads are associated with speed and measures such as speed cameras and the enforcement of regulations have helped to reduce the numbers of accidents: Taylor et al. (2006).

In a comparison of the policies adopted by each county and also the national policies for reducing fatalities by 2010 which were mentioned earlier, it was found that the variations of accidents causing fatal and serious injury over the five-year period match. Accidents have been decreasing in number as these policies have been adopted. With this in mind, the probabilities for other types of accidents are increasing and new policies should focus on them.

REFERENCES

Ayramo, S., P. Pirtala, et al. (2009). Mining road traffic accidents. Software and Computaional Engineering. U. o. Jyvaskyla. Jyvaskyla, Department of Mathematical Information Technology.

Berson, A. and S. J. Smith (1997). Data Warehousing, Data Mining, and Olap. New York, McGraw-Hill.

Beshah, T. and S. Hill (2008). Mining Road Traffic Accident Data to Improve Safety: Role of Road-related Factors on Accident Severity in Ethiopia.

Bioinformatics-and-Information-Mining (2010). "KNIME." from http://www.knime.org/.

Bouckaert, R. R., E. Frank, et al. (2010). WEKA Manual for Version 3-6-2.

Clarke, D. D., R. Forsyth, et al. (1998). "Behavioural factors in accidents at road junctions: The use of a genetic algorithm to extract descriptive rules from police case files." Accident Analysis & Prevention 30(2): 223-234.

Department-for-Transport (2007). Trends in Fatal Car-occupant Accidents. Road Safety Research Report No. 76, Department for Transport.

Department-of-Transport (2001-2006). Road Casualties of Great Britain 2001-2005, Department of Transport.

DETR (2001). Road Accident Data, Variables and Values and Export Record Layouts. T. a. t. R. The Department of the Environment.

Exforsys (2010). "Data mining advantages." Retrieved 30 June 2010, 2010, from http://www.exforsys.com/tutorials/data-mining/data-mining-advantages.html.

Friedman, N., D. Geiger, et al. (1997). "Bayesian Network Classifiers." Machine Learning 29: 131-163.

Gabby, D. M. and P. Smets (1998). Handbook of defeasible reasoning and uncertainty management systems.

GMLTP (2005). Greater Manchester Local Transport Plan. Road Safety Strategy. Manchester.

Heydecker, B. G. and J. Wu (1999). "Identification of sites for road accident remedial work by Bayesian statistical methods: an example of uncertain inference." Advances in Engineering Software 32(10-11): 859-869.

Huebner, R. A. (2008). "Mining for Knowledge in Higher Education." Proceeding of ASBBS 15(1).

Ihaka, R. and R. Gentleman (2010). "R Project." Retrieved 10 July 2010, 2010, from http://www.r-project.org/.

Jensen, T. B. (2009). Intorduction to Bayesian Networks. Copenhagen, University of Copenhagen.

Keim, D. A. (2001). "Visual Exploration of Large Data Sets." Communications od the ACM 44(8).

Link, H., J. S.Dodgson, et al. (1999). The Costs of Road Infrastructure and Congestion in Europe, Springer-Verlag.

Mohan, D. (2002). "Social Cost of Road Traffic Crashes in India1." Transportation Research and Injury Prevention Programme, Indian Institute of Technology.

Pai, C.-J., H.-R. Tyan, et al. (2004). "Pedestrian detection and tracking at crossroads." Pattern Recognition 37(5): 1025-1034.

PIARC (2007). Road accident investigation guidelines for road engineers. World-Road-Associaten.

Reutemann, P. (2010). Percentage Split. Hamilton, University of Waikato.

Taylor, M. C., A. Baruya, et al. (2006). The relationship between speed and accidents on rural single carriageway roads. DfT, Department for Transport. Report number 511.

The-Data-Mine (2010). "All data mining software." Retrieved 23 July 2010, 2010, from http://www.the-data-mine.com/bin/view/Software/AllDataMiningSoftware.

University-of-Waikato (2010). "WEKA." Retrieved 10 July 2010, 2010, from http://www.cs.waikato.ac.nz/~ml/weka/.

Wanvik, P. O. (2009). "Effects of road lighting: An analysis based on Dutch accident statistics 1987-2006." Accident Analysis & Prevention 41(1): 123-128.

WHO (2010). "Road traffic injuries." Retrieved 30/05/2010, 2010, from http://www.who.int/features/factfiles/roadsafety/01_en.html.

Wu, J. and B. G. Heydecker (1993). "A knowledge-based system for road accident remedial work." Computing Systems in Engineering 4(2-3): 337-348.