EXPLORATORY ANALYSIS OF ROAD ACCIDENTS WITH THE SELF-ORGANISING MAP

KONSTA SIRVIO
Sirway Ltd., Finland
konsta.sirvio@sirway.fi

JAAKKO HOLLMÉN
Laboratory of Computer and Information Science, Helsinki University of Technology, Espoo, Finland
jaakko.hollmen@tkk.fi

ABSTRACT

Road accidents are a major concern to the society and for the individuals involved. All the consequences of road accidents motivate the research to understand the factors behind the causalities and the effort to minimise the number of accidents. We investigate the factors behind the road accidents using two databases recording data about the road accidents and the road conditions at the scene of the accidents in Southern Finland. We analyse the data in an exploratory fashion using an artificial neural network called the Self-Organising Map. In addition, we classify fatal accidents using a Naive Bayes classifier. Fatality and injury risks are calculated for various accident conditions. We discuss our findings and map dome directions for further research.

1 INTRODUCTION

Road accidents are a major concern to the society and to the individuals that have experienced one. Besides human deaths, mental and physical injuries as well as financial costs are caused by accidents. All the consequences of road accidents clearly motivate the research to understand the factors behind the causalities and the effort to minimise the number of accidents.

Accidents occur due to various causes such as human error, bad weather and poor road condition. Studies have been conducted about the causes. However, over-encompassing answers have not been found and probably will not be about the reasons, but continuous research might provide with some answers and thereby means for minimising accidents.

In this study, relatively new data-mining methods of computational intelligence are introduced in the field of classification of road accidents and analysis of the risk factors utilising the data in public databases administered by Finnish Road Administration. The utilised methods are Bayes network with strong assumptions about data independency as well as Self-Organising Map (SOM) that belongs to methods of neural networks.

First, an artificial neural network Self-Organising Map (SOM)[1] is applied in the exploratory analysis of the data sets. The SOM reveals the natural clustering structure in the data and is a helpful tool in visualising the cluster structure. Second, we train a Naive Bayes classifier to classify fatal accidents.

[1] Kohonen, T. (1990), pp. 1464-1480.

2 DESCRIPTION OF THE DATA

Finnish Road Administration collects road related data that is stored in databases. In this study two of them are exploited, namely the road condition database and road accident database.[2]

Information about road condition measured by various variables is collected in Finland for the whole network. Collection interval depends on the type of condition variables. Most important factors of road condition are collected on yearly bases at least for the most important paved roads. These condition factors are International Roughness Index (IRI) and rutting. Average collection period for surface defection information such as cracking and edge break is three years. Information about structural condition (central deflection) is also collected, but not utilised in this research since it is assumed that they do not have much relationship with road accidents. Table 1 summarises the variables that are taken into consideration from the road condition database.

Table 1. Variables from the road condition database

1 - Road	8 - Days from year start	15 - Narrow vertical cracking
2 - Section	9 - IRI	16 - Broad vertical cracking
3 - Start distance from section start (m)	10 - Rutting	17 - Narrow horizontal cracking
4 - End distance from section start (m)	11 - Max rutting	18 - Broad horizontal cracking
5 - Year	12 - Net cracking	19 - Potholes
6 - Month	13 - Narrow seam cracking	20 - Ravelling
7 - Day	14 - Broad seam cracking	21 - Edge break

Road accident database is compiled from the information the police collects at the scene of road accidents. The database covers descriptive variables about the accident, participant information and general conditions. The variables utilised in this study from the accident database are presented in Table 2.

Table 2. Variables from the road accident database

1 - Road	10 - Dead	19 - Lightness
2 - Section	11 - Injured	20 - Weather
3 - Distance (m)	12 - Accident type	21 - Location
4 - Year	13 - Accident class	22 - Heavy participant involved
5 - Month	14 - Number of participants	23 - Alcohol involved
6 - Day	15 - Speed limit	24 - AADT
7 - Days from year start	16 - Pavement type	25 - AADT of heavy traffic
8 - Weekday	17 - Surface condition	26 - Number of carriageways
9 - Accident hour	18 - Temperature	27 - Road width

In this case study all the road accidents were selected that happened in the Southern Finland in Uusimaa region during the years 1997-2005. The table of selected accidents was the bases where the condition information was added according to the accident location. Location was determined by road and section numbers from both databases. As the condition surveys were conducted so that average value was measured for each 100

[2] www.tiehallinto.fi

metre long segment of each road, these average values were taken for each accident that had distance information from section start points. The condition survey data was not always coherent and many values were missing. When the most important condition information (IRI, rutting) could not be found for an accident location the accident was removed from the data set. Removal was chosen when the nearest measurement location was over 200 metres from the location and when over two years had passed from the last condition survey on the very location or the next condition survey happened over two years after the accident.

After all the pre-processing the data set was reduced to 21 164 accidents out of which 301 were accidents that caused human casualties. In 5 058 accidents of the study injuries were reported. Table 3 summarises how the data set was formed to the final version.

Table 3. The number of accidents, accidents with injuries and fatal accidents

Original number of accidents	**Original number of deadly accidents**	**Original number of accidents with injuries**
22 935	318	5 487
Final number of accidents	**Final number of deadly accidents**	**Final number of accidents with injuries**
21 164	301	5 058

As a summary of the variables they could be classified to the following types:

1. Variables of road condition
2. Variables of weather conditions
3. Variables of participants
4. Descriptive variables of the accidents

3 RESEARCH METHODS

3.1 Pre-processing

The research was conducted in various phases. The first step was to find the original data. It was received in text and Excel files. These were inserted into a database and into separate tables. A small graphical user interface was programmed in C++. The selection process of correct condition value of the road for each accident was realised by Structured Query Language (SQL) from the databases combined with more complex loops and conditions programmed by C++.

Having completed the data tables they were exported from the database into text files and then imported to Matlab –programme that was used in the final analysis. Some missing values were encountered with measured temperature, lightness category, weather conditions and average traffic on the spot. These variables were included as well as the accidents with before-mentioned missing values. Missing values were replaced by zero. All the data was normalised by Equation 1. Then all the variables had zero as mean value and one as standard deviation.

$$Z = \frac{X - \mu_V}{\sigma} \qquad (1)$$

3.2 Self-Organising Map

Self-Organising Map is an artificial neural network[3] trained in unsupervised mode, that is, without any class labels. The SOM reveals the natural structure in the data and visualises the cluster structure in a high-dimensional data with a low-dimensional display.

The training procedures of the SOM are divided into two steps, which are applied in an alternating fashion: first, one sample is taken from the database and a closest neuron (map unit) in the SOM is searched as the closest vector in the input space. Second, the winner neuron and neurons in its topological neighbourhood are updated to be a little bit closer to the sample. This process is repeated thousands of times and the map converges to represent the input data (with suitably chosen training parameters).[4]

The method finds the internal structure of the data without help of external teacher unlike many other neural networks. As the result a map of selected number of neurons is produced. The map can be utilised for data visualisation as well as for visually or computationally finding possible clusters in the data. Natural clusters can be inspected by so-called U-matrix that reveals the distances of individual neurons. [5]

3.3 K-means -clustering

The idea of clustering is to form similar groups out of the input data. The similarity can be measured by two basic criteria – the minimum distance of data points from the cluster centre among the points belonging to the cluster and the maximum distance between separate clusters. The ratio of these variables called within cluster variation $wc(C)$ and between cluster variation $bc(C)$ can be used as the score function measuring the goodness of the clustering.

K-means clustering is basically the same as SOM –algorithm with the exception that a neighbourhood function is not used, but only the winning prototype is being updated. In K-means clustering algorithm *K* number of clusters are either selected or randomly chosen. The cluster centres is the mean value of the data vectors belonging to a cluster. Each data point is assigned to a cluster where the Euclidian distance between the centre and the data point is minimised. The cluster centres are updated after each assignment of a data point and the algorithm is repeated until no further changes occur thus leading to at least a local optimum.[6]

3.4 Naive Bayes classification of fatal accidents

In Naive Bayes classification, the labels of the classes are known; the training of the Naive Bayes classifier is achieved by the choice of the variables and the estimation of class distributions of the classes. In our example, the following two groups of classes are analysed:

1. Fatal accidents and Non-fatal accidents
2. Accidents with injuries and Accidents with no injuries.

---

3 Kohonen, T. (1990), pp. 1464-1480.
4 Haykin, S. (1999), pp. 446-466.
5 Ultsch, A. and Siemon, H. (1990), pp. 305-308.
6 Hand, D; Mannila, H; Smyth, P. (2001), pp. 296-308.

In Naive Bayes classification, the observed variables are assumed conditionally independent given the class information. While the assumption is often violated in practice, the classification works well despite of the shortcoming.

In Bayesian inference a random variable can be classified into categories according to prior information about the results and measured values of dependent variables of the phenomenon. In the full Bayesian model all the combinations of joint probability distribution functions between variables are modelled and taken into consideration in the calculations. Being very laboursome with large problems only some of the joint functions can be modelled and in the other extreme there is a method called Naive Bayes that has strong assumptions about the variables since they are presumed to be independent with each other. Therefore, joint probability distribution functions are not required to be calculated.

Let us denote occurrence of the phenomenon such as a fatal accident by FA and not fatal accident by NFA. Now, the probability of a fatal accident given the measurement data X is presented in Equation 2. [7]

$$P(FA \mid X) = \frac{P(FA)P(X \mid FA)}{P(X)} \tag{2}$$

Here $P(X) = P(X \mid FA)P(FA) + P(X \mid NFA)P(NFA)$ that leads to Equation 3.

$$P(FA \mid X) = \frac{P(FA)P(X \mid FA)}{P(X \mid FA)P(FA) + P(X \mid NFA)P(NFA)} \tag{3}$$

Equation 3 can then be used in fatality or injury risk calculations taking into consideration of all the measured / collected random variable vectors *X* assuming their independence.

4 RESULTS

All the experiments were performed using the SOM toolbox for Matlab software.[8]

First, the combination of accident and condition data was processed by a SOM and a distance map (U-matrix) between the neurons was produced. It revealed possible cluster borders and raised a question whether the accidents that caused deaths and injuries were different from other accidents in some way. By placing those accidents on the map having injuries (black dots) or deaths (red dots) it could be seen that the deadly accidents were grouped closely together while the accidents with injuries were widely spread as depicted on the left of Figure 1.

[7] Hollmén, J. (2000), pp. 16-18.

[8] Vesanto, Juha and Himberg, Johan and Alhoniemi, Esa and Parhankangas, Juha. (2000)

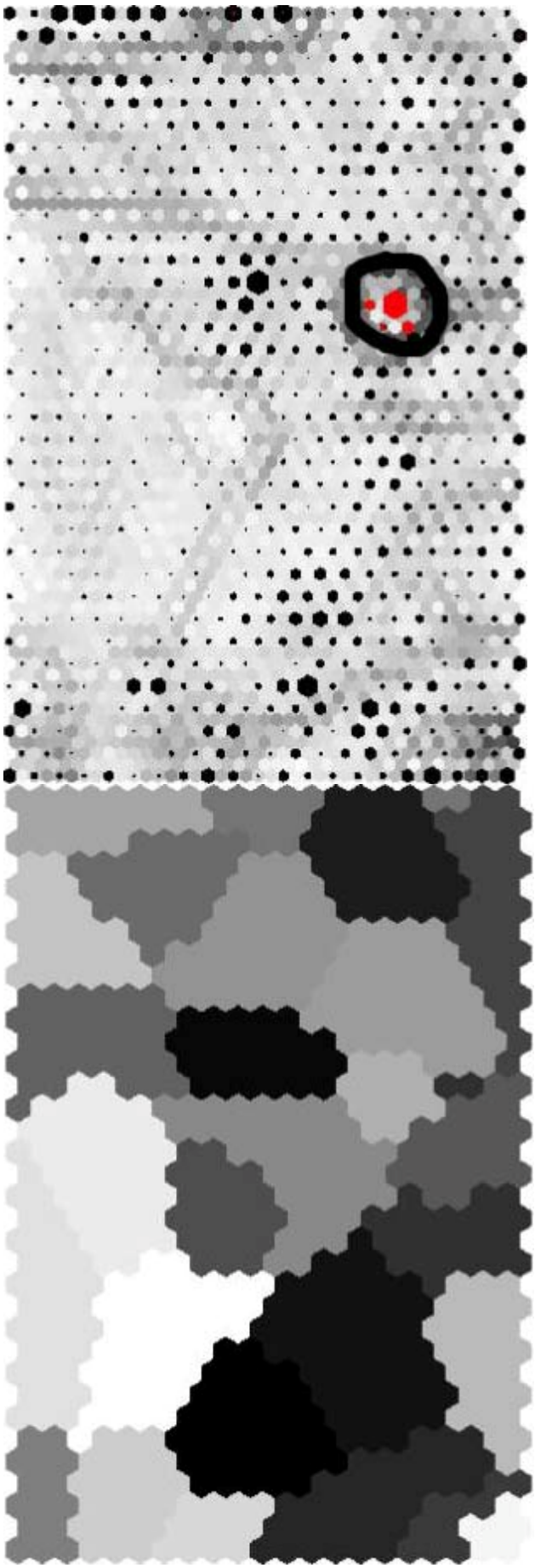

Figure 1. U-matrix of SOM with deadly accidents marked by red dots within the circled area on the left and cluster border on the right

Darker colours in the figure indicate possible borders of clusters. Since the deadly accidents were grouped close to each other, some clustering could be expected. Therefore, the whole set of accidents were clustered in order to see whether the deadly accidents form a uniform group. K-means –clustering was performed to the previously generated map of neurons without restricting the number of clusters. The basis for clustering was distances between the neurons (U-matrix). The clustering algorithm produced 27 clusters out of which one clearly contained most of the deadly accidents. This cluster included 294 accidents and therefore 7 deadly ones were mapped into another cluster.

The normalised variable values for the cluster of deadly accidents were examined by asymmetric representation of the data by plotting the minimum values and the 95% percentile for variable values of deadly accidents. The normalised values were plotted with the same indicators for the whole data set and the results are presented in Figure 2. Here the solid line represents the cluster data and the line with plus-signs the whole data set.

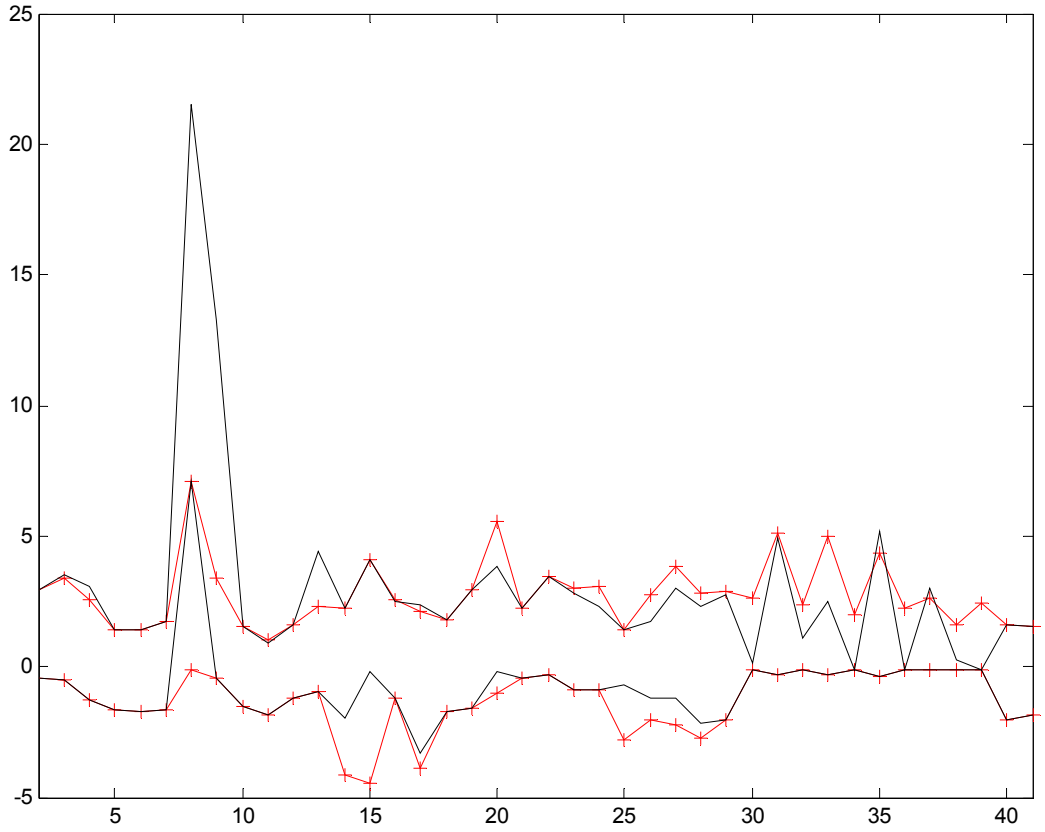

Figure 2. Minimum and 95% -percentile values for the normalised cluster data and for all the data.

Probabilities were calculated for each variable in order to better get the idea of the characteristics of deadly accidents. First, the class probabilities were calculated for the cases: Fatal accident and not fatal accident if an accident happens. Then, the probabilities of different outcomes within these classes were calculated. Naive Bayes –classifier was then applied in order to find out what type of accidents are more probably fatal than not.

As Naive Bayes assumes a strong assumption of having independent variables some variables were excluded having too much correlation with some other variable. Measured rutting and maximum rutting were very much correlated. Surface defects were not

measured uniformly on the network. Accident type gives the same information in more detailed form as accident class. Accident year, road section and distance from the section start were not considered relevant. Number of days from the year start carried essentially the same information as month. Pavement type didn't have much effect on the risk of fatality as well as alcohol. 7.6% of the accidents that caused fatality were driven under influence of alcohol while 7.9% of all the accidents had a driver having consumed alcohol. Majority of accidents happened on the carriageways that were under consideration here. Accident hour didn't seem to have much effect on fatality apart from one peculiarity of those accidents that happened between 15:00 and 16:00. Approximately 8.6% of fatal accidents happened during that time while 5.8% of not fatal accidents occurred at the very same time.

Table 4 summarises the characteristics and conditions of the accidents where the class probabilities favour fatal accidents.

Table 4. Conditional probabilities of variables within different accident categories

	P (X | Fatal accidents)	P (X | Not fatal accidents)	P (X | Injuries)	P (X | Not injuries)
P (category | accident occurred)	**1.42 %**	**98.58 %**	**23.90 %**	**76.10 %**
Road number = 1,2,4,7,25,51,130,148	**51.85 %**	**34.48 %**	30.38 %	36.09 %
Carriageways = 1	72.76 %	67.00 %	64.91 %	67.70 %
80 cm < Road width < 160 cm	**57.15 %**	**47.19 %**	43.91 %	48.41 %
Speed limit => 80	**79.40 %**	**64.85 %**	55.24 %	68.14 %
0.75 < IRI < 1.5	**67.78 %**	**56.47 %**	53.50 %	57.61 %
Rutting < 8	55.48 %	51.74 %	50.53 %	52.19 %
500 < AADT Heavy < 1700	**51.83 %**	**37.43 %**	36.30 %	38.05 %
4000 < AADT < 17000	**52.49 %**	**36.60 %**	35.84 %	37.14 %
Month = March, April or June-September	53.48 %	44.19 %	50.26 %	42.46 %
Weekday = Thursday or Friday	33.22 %	30.98 %	30.72 %	31.10 %
11 < day < 17	20.60 %	16.58 %	16.13 %	16.80 %
Weather = clear	40.86 %	35.69 %	40.43 %	34.30 %
Lightness = daylight	**62.79 %**	**50.52 %**	63.86 %	46.56 %
16 < temperature < 36	26.58 %	17.97 %	24.44 %	16.10 %
Surface = dry and clean	58.14 %	52.50 %	54.84 %	51.88 %
Accident class = meeting of the vehicles	**35.22 %**	**3.33 %**	8.54 %	2.29 %
Heavy participant involved	**38.21 %**	**15.91 %**	11.39 %	16.59 %
Participants > 2	13.63 %	7.74 %	100.00 %	6.71 %
Injured > 0	**45.83 %**	**23.55 %**	15.09 %	0.00 %

The first row informs that if an accident occurs the probability of having casualties is 1.4%. The first group of variables below characterises the road where the accidents occurred. In the second category information of accident time is given. The third group provides with information about the weather conditions. The last group characterises the actual accidents.

Some interesting points can be noted. Most of the fatal accidents in Southern Finland occur on 8 roads. Fatality was increased in narrow roads where the speed limit was highest. Fatality risk seemed to increase also on the roads with low roughness (IRI) and rutting e.g. roads in good condition. This occurred also on the spots of having medium Average Annual Daily Traffic (AADT) both for all the traffic as well as for the heavy traffic. Some variation can be observed with temporal risk of fatality as deadly accidents happened more probably just around the middle of the month than other accidents. Also, spring, summer and early autumn had most of the fatal accidents.

In accident prevention the priority should be set so that those are tackled first where the probability of fatality is higher than non-fatality. By utilising Equation 3 and replacing $P(FA)$ by 0.0142, $P(NFA)$ by 0.9858 and $P(X \mid FA)$ as well as $P(X \mid NFA)$ by the probabilities calculated in Table 1 fatality risks can be calculated. When various conditions are met, the conditional probabilities $P(X \mid FA)$ and $P(X \mid NFA)$ are products of probabilities of individual conditions according to the strong assumption of independence of the variables.

The division of accidents to other classes in fatal and not fatal accidents as well as in accidents with injuries and not injuries are presented in Table 5.

Table 5. Conditional probabilities of various accident classes

Accident class	Fatal	Not fatal	Injuries	Not injuries
Individual accident	26.58 %	25.46 %	34.58 %	22.62 %
Turning accident	3.32 %	4.67 %	7.75 %	3.68 %
Overtaking accident	3.65 %	5.81 %	4.17 %	6.28 %
Crossing accident	6.31 %	6.61 %	11.31 %	5.13 %
Meeting of the vehicles	35.22 %	3.33 %	8.54 %	2.29 %
Rear-end collision	2.33 %	10.62 %	12.69 %	9.81 %
Motorcycle accident	1.33 %	1.18 %	3.28 %	0.53 %
Bicycle accident	5.65 %	1.69 %	5.40 %	0.60 %
Pedestrian accident	10.3 %	1.02 %	3.54 %	0.40 %
Moose accident	2.66 %	10.98 %	4.63 %	12.82 %
Deer accident	0.00 %	23.53 %	0.67 %	30.27 %
Other animal accident	0.00 %	1.12 %	0.38 %	1.33 %
Other accident	2.66 %	3.97 %	3.06 %	4.23 %

In Table 6 an example is shown how fatality and injury risks increase in an accident having increasing number of conditions met.

Table 6. Fatality and injury risks under various conditions

Condition	Fatality risk	Risk of injury
Category probability	1.42 %	23.90 %
& Accident class: Meeting of the vehicles	13.22 %	53.96 %
& Heavy participant involved	26.79 %	51.59 %
& More than 2 participants	39.18 %	64.40 %
& Road number: 1,2,4,7,25,51,130 or 148	49.21 %	60.37 %
& IRI € [0.75, 1.5]	53.77 %	58.58 %
& AADT € [4 000, 17 000]	62.52 %	57.43 %
& Lightness = daylight	67.46 %	64.92 %

When the 5 first conditions of Table 6 and at least one other condition from Table 5 are met, Casualties are more probable in an accident than mere injuries or damage to the property. It is also interesting to notice that the injury risk remains more stable, but over 50% in the accident class of meeting of the vehicles.

5 DISCUSSION

The results show that fatality risk is increased most when a road accident is either pedestrian or meeting of the vehicles. This is natural having usually highest relative speed between the vehicles in collisions and in the second case pedestrians do not usually have

much protection for masses over 1 000 kg moving in higher relative velocity. And how could the risk be reduced? One possibility would be by having railings separating lanes of the opposite driving directions or constructing separate carriageways. Relatively small number of accidents was classified as overtaking accidents and much higher as meeting of the vehicles. Either the investigations did not classify them right or drivers collided due to other reasons. There is a plethora of reasons. They can be suicides, falling asleep, fit, slippery road or reading emails and short messages from a mobile phone.

As the accidents examined happened on the network of the Finnish Road Administration they were mostly on rural areas. As pedestrian accidents were present it can be assumed that the casualties were walking on the road side and light traffic lanes did not exist. A simple solution is to build light traffic lanes along with roads when light traffic is expected.

Accidents with heavy traffic naturally increase the risk of death. No simple solution exists, but various methods ought to be used such as favouring transport of goods on the rail network, better control of driving times and conditions of the vehicles or separate lanes.

When accidents had more than 2 participants the fatality risk increased greatly. Also, the fatality risk was increased under the conditions where roughness measured by IRI was between 0.75 and 1.5, rutting below 8 and when the traffic measure AADT was between 4 000 and 17 000. However, under the same conditions injury risk was decreased. Good road conditions may give a false impression of safety thus causing over speeding this explanation is supported by having higher fatality risk on daylight with dry and clean surface during the summer time when the weather is clear and temperature between 16 and 36 degrees in centigrade. However, keeping roads in lower condition would increase the road user costs.

Relatively narrow roads (80-160 cm) and existence of only one carriageway increased the fatality risk. Dual carriageways would reduce fatality risk and collisions of the vehicles.

According to the results of the study we suggest road administrations to take the following actions on the riskiest roads whenever feasible:

1. Separation of the possibility of meeting of the vehicles of the opposite direction by separate carriageways or extra lanes with railings in the centre
2. Widening of the risky roads
3. Reduction of lorry traffic with lighter traffic by a lane for the heavy traffic, transfer of the freight traffic to the rails and stricter control over vehicle and driver conditions
4. Separate light traffic lanes whenever possible
5. Traffic safety campaigns scheduled to the spring and summer time

6 FUTURE RESEARCH

Due to the complexity of the phenomenon more data could be added into the analysis in order to find more accident causes and increase reliability of the risk probabilities. Some of the data can be acquired rather easily. This includes variables of road geometry including variables such as number of percentage of rises and falls, horizontal curvature and elevation. Road geometry affects visibility and handling of a vehicle and could thus be a cause for some accidents. The accident database contains in a separate table some more information about the participants such as age and sex. With some work these could be included in the future research. Other data items could be historical weather measurements.

An interesting comparative research would be to take separate regions of Finland and investigate whether there are differences in accident causes and for what reasons. The whole data of one country could be grouped, similar analysis conducted and then compared with other countries with similarly collected data. This could reveal best practices in prevention of accidents along various countries.

More accurate methods can always be developed and introduced in the study. Here, a method of Naive Bayes was utilised. The assumption of having independency between the variables might be too strong and therefore real Bayes network could be utilised when strong dependence between variables exist if the required computational power is not a hindrance.

ACKNOWLEDGEMENTS

We would like to express our gratitude to Ismo Iso-Heiniemi, Raija Huhtala and Markku Visti from the Finnish Road Administration for providing the original data. Besides, the utmost gratitude is addressed to the Finnish Cultural Foundation for a grant for the doctor studies. In addition to above mentioned parties, we would like to thank every one that has helped in the research process.

REFERENCES

1. Kohonen, Teuvo. (1990). The Self-Organizing Map. Proceedings of the IEEE volume 78 number 9
2. www.tiehallinto.fi
3. Kohonen, Teuvo. (1990).
4. Haykin, S. (1999). Neural Networks. A comprehensive foundation, Prentice-Hall Inc. London, United Kingdom
5. Ultsch, A. and Siemon, H. (1990). Kohonen's Self-Organizing Maps for exploratory data analysis. Proceedings of the International Neural Network Conference (INNC'90). Kluwer
6. Hand, D and Mannila, H and Smyth, P. (2001). Principles of Data Mining. MIT Press. Cambridge, United States of America
7. Hollmén, Jaakko. (2000). User profiling and classification for fraud detection in mobile communications networks. Helsinki University of Technology. Espoo, Finland
8. Vesanto, Juha and Himberg, Johan and Alhoniemi, Esa and Parhankangas, Juha. (2000). SOM Toolbox for Matlab 5. Helsinki University of Technology, Laboratory of Computer and Information Science. Espoo, Finland (Software freely available via WWW at URL: http://www.cis.hut.fi/projects/somtoolbox/)

.