Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 160 (2014) 607 – 614

Procedia
Social and Behavioral Sciences

XI Congreso de Ingeniería del Transporte (CIT 2014)

Using data mining techniques to road safety improvement in Spanish roads.

Luis Martín, Leticia Baena, Laura Garach, Griselda López and Juan de Oña*

TRYSE Research Group, Department of Civil Engineering, University of Granada, ETSI Caminos, Canales y Puertos, c/ Severo Ochoa, s/n, 18071 Granada (Spain)

Abstract

Crashes are events that involve the interaction of different components: road, driver, vehicle and environment. Nevertheless, road is an essential component and improvements on road conditions are directly related to increased traffic safety. From 2008 to 2010 a road safety inspection project was developed, whose aim was to identify and collect information about hazardous points on the Complementary Road Network of Andalusia, Spain, and build a database with this information. These elements were technically called Susceptible Elements of Improvement (ESM), which are defined as elements on the road that show worse road conditions than the ideal road safety standards.

The main objective of this paper is to study the relationship between ESMs, number of crashes and hazardous sections, by analysing the information gathered in this database with advanced data mining techniques.

Economically, this project is rather beneficial, since the resources of governments are limited, and therefore, it is necessary to intervene in those sections that have a higher cost-effectiveness ratio. Therefore, these relationships between roads conditions and crashes will be identified by analysing the information available in this data set of the Government of Andalusia, which has not been previously used.

Keywords: road safety; crashes; hazardous sections; data mining.

* Corresponding author. Tel.: +34 958 24 99 79; fax: +34 958 24 99 79.
E-mail address: jdona@ugr.es

doi:10.1016/j.sbspro.2014.12.174

1. Introduction

The accident rate is a problem that exists for the last century, due to the increment of the traffic rate. According to a report by the World Health Organization, 50 million injured and 1.27 million deaths a year are caused by crashes. For that reason road accidents are considered to be a worldwide problem.

Moreover, crashes involve governments in huge costs both economic and resources. They are a complex phenomenon that implicates the interaction of different elements: road, driver, vehicle and environment. Nevertheless, roads are an essential component and an improvement in road elements would enhance in safety roads.

In 2008, the Andalusia Regional Government hired a road safety inspection of the Complementary Road Network of Andalusia (Spain). The outcome of this study was a database that identifies all elements of the road or its environment that cause a deficit in road safety. Those items are named Susceptible Elements for Improving (ESM).

Given the potential of this information, in order to improve road safety, the authors proposed a research project whose aim is to analyse the Andalusia Complementary Road Network, by relating ESM, crashes, and hazardous sections (TCA), which are sections with potential for improving road safety, using advanced data mining techniques.

During the last years, several studies about crashes have been carried out, using regression models. These models need fixed hypothesis and predefine a relationship between dependent and independent variables. Erroneous estimations could be produced if hypothesis are not fulfilled in those models (Chang and Wang, 2006). In addition, hidden knowledge that exists on databases cannot be extracted by those traditional methodologies.

In order to solve these deficiencies, in the last decade, several data mining techniques have been applied on the field of road safety (Sohn and Shin, 2001; Chang and Wang, 2006; Kashani and Mohaymany, 2011; Kashani et al., 2011). Although several road safety researchers have applied these techniques, in Spain the TRYSE research group has been the solely group in applying them in this field. They have studied crash-injury severity with very successful outcomes (De Oña, Mujalli & Calvo, 2011; Mujalli & De Oña, 2011; De Oña, López & Abellán, 2013; De Oña, López, Mujalli & Calvo, 2013).

The objectives of the research is to build a unique data bank that contains the three extensive databases (ESM, roads and accidents), to identify hazardous sections in the studied roads and to apply advanced data mining techniques in order to discover hidden relationships between characteristic of the roads, ESM and crashes.

The paper is organized in three mayor sections. Section 1 presents a brief introduction, Section 2 comprehends the data and methodology and Section 3 presents the summary of the study.

2. Data and methodology

2.1. Data

Three databases are need for the present study:

- Database of ESM: it has been provided by the Andalusia Regional Government. These elements exist on the road if there is a deviation from the ideal conditions from the viewpoint of road safety.
- Database of roads: it has been provided by the Andalusia Regional Government, and it is an inventory of the Complementary Road Network of Andalusia. It has information that describes the characteristic of roads, such as geometry or equipment.
- Database of crashes: it has been provided by the Spanish National Government (Dirección General de Tráfico, DGT) and it contains information of all crashes happened between 2006 and 2008.

2.1.1. Susceptible Elements for Improvement

Susceptible Elements for Improving (ESM) exists on a road when a deviation from the ideal conditions is produced, from the viewpoint of road safety. For instance, a road access must be signposted in order to advise the driver how manoeuvre must be done. Whether the signal is not placed or was incorrectly placed, it will be a vertical signalling ESM.

Eight families of ESM were identified:

- Layout
- Signalling and marking
- Containment systems and road obstacles
- Intersections, junctions and roundabouts
- Crossing
- Tunnels
- Road access
- Others circumstances

Moreover, each family was subdivided into different ESMs' groups. Thus, the elements are perfectly classified in order to analyse their influence on road safety.

The existence of these elements was identified creating a database with all the shortcomings of the roads of Andalusia at the time of the inspection. ESMs may exist along stretches or in sections. For example, within the family "signage and marking", the element 2101 "Unreadable vertical signal" would be given in a section where the signal is placed. It would be checked if a driver travelling on the road could read the signal. If it is not readable, as in the example of Figure 1, an ESM was identified on the section.

This database collected all the ESM located in each road section, or point, in which an ESM exists.

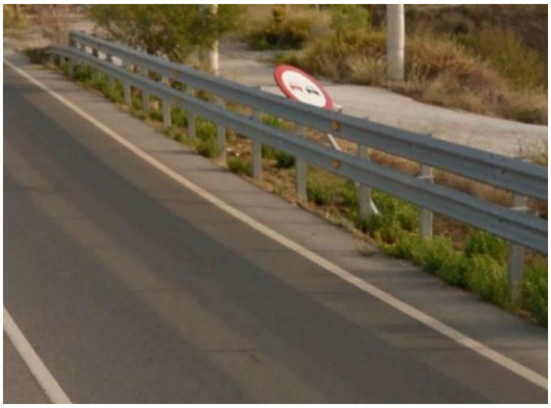

Fig 1. ESM 2101: Unreadable vertical signal

2.1.2. Data of roads

The data collected at the base describe characteristics of road. Some of these data are: radio, cant, slope, curve length, etc. These data are recorded every 10 meters, all over the Andalusia Complementary Road Network.

2.1.3. Data of crashes

In order to analyse crashes, it is need to define a period of time of the study. Owing to the fact that the aim of the present project is to correlate crashes and ESMs, the period of time must be the period of inspection. So, we will consider crashes from 2006 to 2008.

These data have been provided by the DGT. Nevertheless, they supplied information about all the crashes happened in all the roads of Andalusia. Therefore, the first step is to filter them, and obtain the accidents on the roads under consideration for this project (the Complementary Road Network of Andalusia).

The information collected in this database describes the features of the crashes, such as location, the number of vehicles involved, the type of accident, etc. Crashes are located in 100 metres long sections.

2.2. Methodology

The methodology is as follow. Firstly, it is necessary to build a data bank that contains the three extensive databases (ESM, roads and accidents). It should be verified, due to the fact that the reliability of the outcomes depends on the reliability and correspondence of the databases.

Secondly, hazardous sections must be calculated. Spanish methodology will be used, which take into account annual average daily traffic (AADT) and the number of crashes of each road.

Finally, advanced data mining techniques will be applied in order to discover hidden relationships between characteristic of the roads, ESM and crashes

2.2.1. Database integration

These three databases are not organised in the same way and, for this reason, it is necessary to do a preprocessing step in order to unify the three databases. It is also necessary to handle with outliers and eliminate irrelevant features that could produce a data deviation.

On the other hand, in order to join the three databases, it is necessary to find a common field to relate all data. In road database there is a common data between it and ESM database: distance from the origin. Nevertheless, this data (distance from the origin) is not in the accidents database. So, in order to join these two databases, it is used the field Pk.

2.2.2. *Identification of hazardous sections*

In literature, different definitions and approaches to the identification of hazardous sections could be found. Thus, in some countries, a hazardous section is identified by comparing to a normal level of security; some researches consider recorded data while others consider estimated data. Some definitions of black spots consider accident severity; other definitions do not (Elvik, 2008).

In this project, the Spanish methodology will be used. This methodology is also used by the Andalusia Regional Government as well as by other road administrations.

In Spain, a hazardous road location is a 1 km section in which both the number of injury accident in a five years period and the medium hazardousness index are higher than the average of similar sections.

The variables that establish similar road sections are:

- Typology of road.
- Daily traffic volume.
- Proximity to urban zones.

If any appreciable changes have been performed in the last five years, it will be considered the medium hazard index and the crashes during the time the road section is on its current configuration.

Once road and crashes variables are collected, in order to identify hazardous sections, it is necessary calculate the parameters below:

- IPM5: Medium hazard index for the last five years (acv/10^8veh-km).
- IPM2: Medium hazard index for the last two years (acv/10^8veh-km).
- IPM_{aa}: Medium hazard index for the two years back (acv/10^8veh-km).
- IPM_{ua}: Medium hazard index for the last year (acv/10^8veh-km).
- Σ ACV5: Sum of the crashes with personal injuries for the last five years.
- Σ ACV2: Sum of the crashes with personal injuries for the last two years.
- Σ ACV_{aa}: Sum of the crashes with personal injuries for the two years back.
- Σ ACV_{ua}: Sum of the crashes with personal injuries for the last years.

IPM is calculated using the formula:

$$\mathrm{IPM}=\frac{\mathrm{N}.10^8}{\mathrm{t.V.L}} \qquad (1)$$

Where

- IPM: Medium hazard index per 100 millions of vehicles per kilometre for the period of time consider.

- N: Sum of crashes with personal injuries in the road section analysed and for the period of time considered.
- t: period of time considered (years).
- V: mean of daily traffic volume of the section analysed for the period of time considered.
- L: length (kilometres).

In Spain, a road section is a hazardous section if it satisfies equation 2 and at least one of the criteria below

$$IPM5 \geq P \text{ and } \sum ACV5 \geq N \qquad (2)$$

- Criterion 1: Hazard index in both last 2 years is greater or equal than P/2. It is $IPM_{aa} \geq P/2$ and $IPM_{ua} \geq P/2$.
- Criterion 2: Medium hazard index in the last 2 years is greater or equal than 2P/3. It is It is $IPM2 \geq 2.P/3$.
- Criterion 3: The sum of injury accidents in both last 2 years is greater o equal than N/5. It is $\sum ACV_{aa} \geq N/5$ and $\sum ACV_{ua} \geq N/5$.
- Criterion 4: The sum of injury accidents in the last 2 years is greater o equal than N/2. It is $\sum ACV2 \geq N/2$.

N and P are two thresholds depending on the type of section (type of the road, area, traffic). P has been calculated, using levels of hazard of all sections with similar characteristics, depending on the sum of the average of the series and its standard deviation. N has been calculated, taking into account the number of fatalities of all sections with similar characteristics, depending on the sum of the average of the series and their average deviation. Sections of 1 km may not be coincident with the passing points on the road, and in the case of identified several overlapping TCA, its study will be done jointly, which will lead to the study of over 1 km long sections. Both are obtained Spanish Road Safety Authorities through a statistic study in road of homogenous characteristic. These thresholds are actualised annually.

Table 1 summarizes each criterion (Ministerio de Fomento, 2008):

Table 1. Conditions for calculating TCA.

$IPaa \geq P/2$ and $IPua \geq P/2$	Criterion I
$IPM2 \geq 2P/3$	Criterion II
$\Sigma ACVaa \geq N/5$ and $\Sigma ACVua \geq N/5$	Criterion III
$\Sigma ACV2 \geq N/2$	Criterion IV

P and N values have been calculated by the Spanish National Government. It could be found in Ministerio de Fomento (2008).

In this project P and N values calculated by Andalusia Regional Government would be used. These values have been estimated for the specific case of Andalusia, due to the special characteristics of Andalusia Road Network.

Table 2. Values of P and N.

AADT*	P	N
0 - 500	72	5
500 - 3.000	239	6
3.000 - 5.000	147	5
5.000 – 10.000	100	6
> 10.000	64	8

* The values of P and N are to be considered for each direction. Otherwise, they must be multiplied by 2. Nevertheless, the AADT is for both directions, corresponding to the last year of the study period.

2.2.3. Application of data mining techniques

Finally, data mining techniques are applied in order to discover hidden relationships between characteristic of the road, ESM and crashes. Data mining is non-trivial extraction of implicit, previously unknown and potentially useful information from data. Its aim is to explore and analyse, by automatic or semiautomatic means, large quantities of data in order to discover meaningful patterns. Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al., 1996).

There are different techniques that allow the extraction of unknown information from data. Overall, a technique is the conceptual approach that allows extraction and a given algorithm implements it.

The data mining techniques could be classified into two groups depending on the process used for the extraction of knowledge: supervised and unsupervised techniques (Weiss and Indurkaya, 1998).

- Supervised or predictive techniques. They are characterised by learning from examples. Input examples are accompanied by a class or correct output. So there is a special attribute, usually called class, present in all cases, which specifies whether a case belongs to a certain class, which will be the focus of learning. It is subdivided into classification techniques or regression.
- Unsupervised or descriptive techniques. They learn by observation. There is not a special attribute to guide the learning process. Descriptions, hypotheses or theories are constructed from a dataset without a prior classification of the examples. It is subdivided in Clustering and Association Rules.

Figure 2 shows the main data mining techniques. All of them enable to obtain hidden knowledge. The final choice of the technique mainly depends on the information in the database and the model that is going to be used. This project tries to extract some behaviour patterns from road accidents.

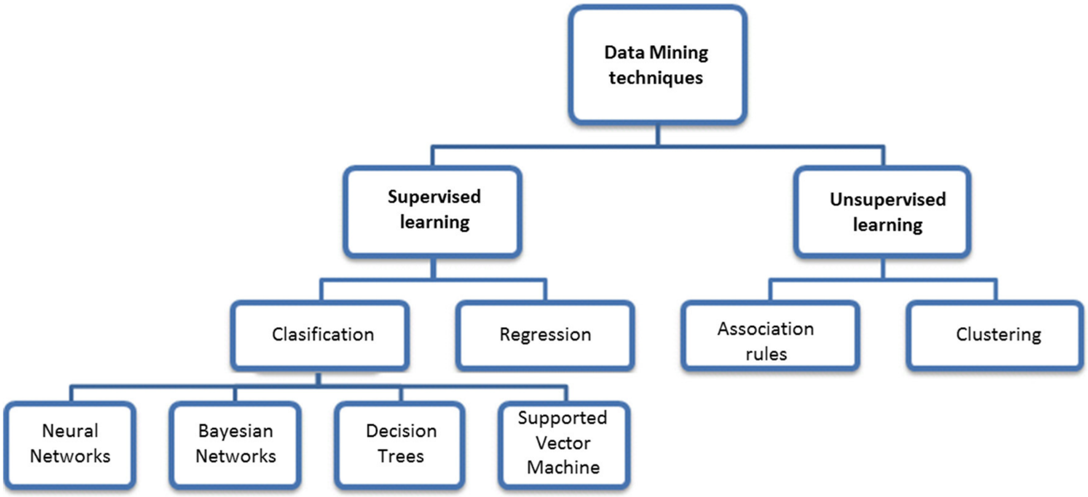

Fig 2. Main data mining techniques.

Several data mining techniques enable to obtain decision rules (IF-THEN) (Kashani et al., 2011), that can help to understand the accident occurrence and the factors involved on it. These techniques are highly interpretable and could be used to expert learning. These techniques are association rules and decision trees.

Therefore, in this project data mining techniques will be used in order to correlate ESM, crashes and TCA. Particularly, decision tree will be applied with the aim of obtaining relations like (IF-THEN) that are easily understandable.

The authorities may use these rules to implement preventive measures, as well as corrective measures, in those road sections that require them. Those improvements will have implications both from the point of view of the reduction in the number of accidents and its severity. Thereby, this project will exploit all the valuable information that the Andalusia Regional Government has about their Complementary Road Network.

3. Summary

This project aims to improve road safety in the Andalusia Complementary Road Network. Thus, it is intended to apply data mining techniques that will enable to identify characteristics of the roads, defined through the elements susceptible of improvement (ESM), which have greater influence in crashes or, in particular, in hazardous sections (TCA).

Once road factors related to accidents are known, the road safety authorities could implement preventive and corrective measures in sections of roads that require them. It should be noted that, from an economic point of view, this project is extremely beneficial, since the resources of the Government are limited, and the cost-effectiveness ratio is expected to be highly important.

As result of this project, the Andalusia Regional Government will have an innovative methodology in terms of:

- A methodology for the analysis and the extraction of information about ESM of its road network. Novel data mining techniques will be used in order to innovate on the study of crashes, an area that has been traditionally focused by conventional statistical techniques. In addition, data mining will allow the identification and the extraction of patterns of behaviour in accidents, which can be explained in a simple and easily understandable way by certain rules. These rules can be used later, for the training of expert systems.
- The interrelation of the ESM of the road, obtained from road safety audits; road crashes and TCAs, enables the identification of hazardous factors of the road. This will let the Administration develop a plan to correct the most dangerous factors identified.
- Road safety inspections are relatively novel, and they are based on filling checklists. This project would be the first study that analyses data from these inspections, crashes and TCAs, simultaneously. Therefore, it would place the Andalusia Regional Government at the international road safety forefront of R&D.

The results will be obtained as rules where certain factors, such as the main causes of accidents, will appear. Those ESM more related to crashes or hazardous locations should be prioritized in order to reduce the accident rates. These results may also be taken into account when new infrastructures will be planed.

Acknowledgements

The authors would like to thank the FEDER of European Union for financial support via project “Análisis de la relación entre Elementos Susceptibles de Mejora, Accidentes y TCA” of the “Programa Operativo FEDER de Andalucía 2007-2013”. We also thank all Agency of Public Works and the Ministry of Furtherance and Housing of Andalusia Regional Government staff and researchers for their dedication and professionalism. Griselda López Maldonado acknowledges to Consejería de Economía, Innovación y Ciencia of Junta de Andalucía for “Beca de formación de personal docente e investigador en Áreas Deficitarias”.

References

Chang, L.Y. and Wang, H.W. (2006). “Analysis of traffic injury severity: an application of non-parametric classification tree techniques”. Accident Analysis and Prevention 38, pp. 1019–1027.

De Oña,J., Mujalli, R. and Calvo, F. (2011). Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. Accident Analysis and Prevention, 43, pp. 402-411.

De Oña, J., López, G. and Abellán, J. (2013). Extracting decision rules from police accident reports through decision trees. Accident Analysis and Prevention, 50, pp 1151-1160.

De Oña, J., López, G., Mujalli, R. and Calvo, F. (2013). Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. Accident Analysis and Prevention, 51, pp 1-10.

Elvik, R., (1997). Evaluations of Road Accident Blackspot Treatment: A Case of the Iron Law of Evaluations Studies? Accident Analysis and Prevention, 29, pp 191-199

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. (1996). “From Data Mining to Knowledge Discovery: An Overview”. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining. AAAI Press pp. 1–34.

Kashani, A., Mohaymany, A., (2011). Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models.

Safety Science 49, pp. 1314-1320.

Kashani, A., Mohaymany, A., Ranjbari, A., (2011). A Data Mining Approach to Identify Key Factors of Traffic Injury Severity. Promet-Traffic & Transportation, 23 (1), pp 11-17.

Ministerio de Fomento (2008). "Nota de Servicio julio 2008. Programa de Seguridad Vial 2009-2011."

Ministerio del Interior, (2009). Las principales cifras de la Siniestralidad Vial. http://www.dgt.es/was6/portal/contenidos/es/seguridad_vial/estadistica/publicaciones/princip_cifras_siniestral/cifras_siniestralidadl009.pdf

Ministerio del Interior, (2011). Estrategia de Seguridad Vial 2011-2020. http://www.dgt.es/was6/portal/contenidos/documentos/seguridad_vial/planes_seg_vial/estrategico_seg_vial/estrategico_2020_004.pdf

Mujalli, R. and De Oña, J. (2011). A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. Journal of Safety Research, 42, pp 317-326.

Servicio de Conservación y Dominio Público Viario. "Documento de Síntesis. Programa de Seguridad Vial 2011". Dirección General de Infraestructuras. Junta de Andalucía.

Sohn, S.Y. and Shin, H.W., (2001). Data mining for road traffic accident type classification. Ergonomics 44, 107–117.

Weiss, S. M. and Indurkhya, N., (1998). Predictive Data Mining. Morgan Kaufmann Publishers.

World Health Organization (2009). “Informe Global sobre el estado de la Seguridad Vial: Tiempo para la Acción”. www.who.int/violence_injury_prevention/road_safety_status/2009