A STUDY OF TRAFFIC ACCIDENTS IN SPANISH INTERCITY ROADS BY MEANS OF FEATURE VECTORS

D. ÚBEDA, A. GIL, L. PAYÁ & O. REINOSO
Department of Systems Engineering and Automation, University Miguel Hernández de Elche, Spain.

ABSTRACT

Frequently, road traffic accidents are modelled as discrete and independent random and rare events, which possess a low probability of occurrence through time. Nevertheless, in order to study each accident individually it is necessary to obtain details of a number of characteristics that surround it, which may be correlated with each other. In this article, we propose to associate the probability of occurrence of an accident with a large number of features such as weather conditions, incidents caused by the start and end of a roadwork, geographical location of speed control radars, roadway infrastructure, etc. The influence of these features is significant and should be taken into account when proposing measures to help alleviate these undesirable events. The big data methods employed to extract the variables or features allow us to compose a series of vectors that will serve as a basis to study road accident distributions.

Keywords: road traffic accident, road traffic data mining, weather features vectors.

1 INTRODUCTION

Road accidents are the cause of a large number of human losses in the world and also imply significant psychological and physical trauma. In addition, according to the European Union [1], the annual economic burden of road accidents in Europe is estimated to be between EUR 10 and 14 billion and the cost of a single fatality may be as high as EUR 1 million.

In this article, we consider road accidents as rare events: that is, the frequency of their occurrence is low. In order to be able to study them, we propose a multivariable system. Their many characteristics were obtained by big data techniques and the entire system was studied using data mining algorithms. Big data methods are employed to extract the variables or features that allow us to compose a feature vector that can be associated to each accident. These vectors are formed by a set of variables that were involved in the accident. These variables include weather conditions, traffic sensors and more and serve as a basis to study road accident distributions on them.

1.1 Studying accidents as rare but massive data events

The treatment of rare events, that is, those that occur with a low probability, is a complex issue. The rare event law or Poisson process consists of a stochastic process defined in a continuous time and is substantiated by counting rare events that happen during that time. The time between each pair of consecutive events possesses an exponential probability distribution with a λ parameter, and differences between different pairs of consecutive events are considered independent.

Recently, these kind of problems based on Poisson processes have been tackled by means of data analysis and data modelling techniques derived from artificial intelligence. These events are treated as massive data (although infrequent) due to the high quantity of variables involved in each one of

WIT CONFERENCES
This paper is part of the Proceedings of the International Conference on Big Data (Big Data 2016)
www.witconferences.com

ISSN: 1755-7437 (paper format), ISSN: 1755-7445 (online), http://www.witpress.com/journals
DOI: 10.2495/DNE-V11-N3-317-327

them. One could imagine, for example, that the probability of occurrence of an accident is greater on a rainy day than on a sunny day. In addition, it could be reasonable to think that the risk of suffering an accident is greater during rush hours compared with the situation when only few vehicles are circulating. As a result, there is a need to store a high quantity of data associated to each accident in order to be able to study them later and try to find important correlations between them. Consequently, big data techniques need to be applied to the problem, since the problem requires storing a great quantity of information associated to each event.

In this article, in addition to the features mentioned above, other features have also been analysed, whose influence on the accident has been deemed to be useful to include. The application of these new features, which, as far as the authors know, have not been employed before in the analysis of traffic accidents is the main contribution of this article. One of the most relevant variables is the impact of the proximity of fixed speed limit radars on the driver's behaviour, or the circulation of vehicles through a section under roadwork in a certain route.

2 FEATURE SELECTION AND OPEN DATA IN SPAIN

2.1 Factors that influence the probability of occurrence of road accidents

During the past years, road accidents have been analysed by different researchers with the objective of establishing models that will allow finding answers to the causes that resulted in each accident, as well as finding solutions that may alleviate the frequency of these undesired events. Therefore, typically, authors have selected features that could be considered to have a larger impact on the frequency or occurrence of road incidents. In most cases, variables relating to weather, characteristics of the road or driver conditions have been used.

Driving under the influence of alcohol or drugs is unquestionably dangerous. We call it "*Classic Road Accident Analysis*", as illustrated in Figure 1. The most relevant features studied during the past years are listed in the subsequent paragraphs.

2.1.1 Driver's behaviour features

Attending to the literature, the authors have typically associated some features to the accident. Sabey and Taylor [2] used car speed at the time of the accident, whereas, Hakim *et al.* [3] and Wang *et al.*

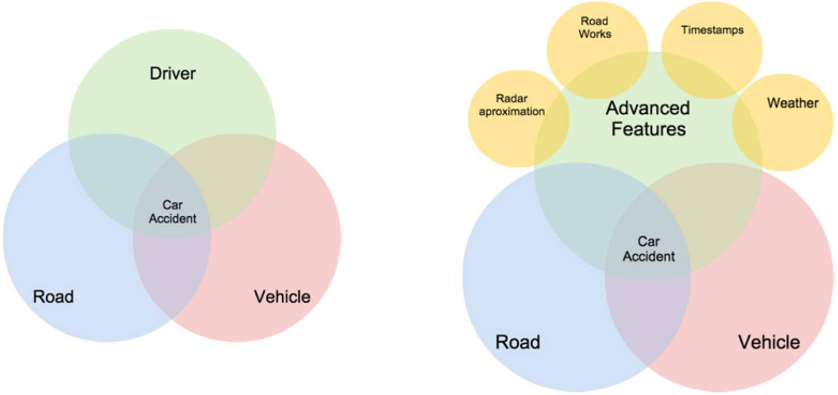


Figure 1: Classic road accident analysis vs. advanced road accident analysis.

[4] studied the speed limit of the road. The use of safety measures was a variable in Simoncic's [5] research, among others. Driver classification was used in other studies as one of the main features: age was used by Davison [6], and experience was also studied by Simoncic. Over the years, these features have evolved and other new features have also come upon, such as the use of mobile phones or other electronic devices when driving.

2.1.2 Features related to the design of the roadway

To date, the characteristics related to the design of the road have been considered to be very important features compared to other possible variables involved in an accident. This can be explained as to why causes of accidents have been historically sought, which has originated a true interest for improving road infrastructure, thus reducing the accident rate.

In the current state of knowledge of the features, it is possible to highlight some of the following:

- Lane width and Average Daily Traffic expressed in veh./day (Zegeer *et al.* [7])
- Width of shoulder of the road (Kalokota *et al.* [8])
- Number of lanes (Wang *et al.* [4])
- Section length (Shankar *et al.* [9])

2.1.3 Weather patterns

Extensive research has correlated weather variables such as rain, fog or snow, with the cause of the occurrence of accidents; however, this is always approached from the point of view of user *mobility*. It is important to identify weather situations that increase or reduce mobility. For instance, inclement weather reduces mobility, so it is important to consider temperature as an appropriate feature in mobility.

In addition, mobility is strongly affected by weather conditions and also influences the number of road accidents and casualties significantly. The causes of the accidents also depend on the type of road (motorways, rural roads or urban roads). Moreover, as said by others, since the weather affects mobility, there is a direct correlation between weather conditions and the number of injury accidents and casualties.

A two-stage approach has been adopted to take account of the climatic factors: first, we seek variables that were open access and then find a correlation of these features with the occurrence of accidents.

2.2 Current situation of open data related to road traffic in Spain

2.2.1 Accident datasets in Spain

Currently, the Spanish Traffic Department (DGT) publishes in its 'open data historical statistical website (2008–2013)' [10] data related to accidents, roads, injured people as well as information concerning each vehicle involved.

As can be seen from Tables 1–3, the Traffic Department publishes numerous variables related to the accident, vehicle, and injured people. These features can help us understand some of the causes involved in an accident. However, we believe that there is a lot of important information missing, which can undoubtedly help researchers improve their studies. For instance, the accurate geographical location of each accident is not provided. Lane information is also missing.

2.2.2 Creating a database suited for our specific needs

As mentioned in the previous paragraph, locating the accident in order to derive other features from its situation is of uttermost importance. In order to address this, we started to study alternatives to the Spanish statistical website.

Table 1: DGT public features of an accident.

Field names		
Accident ID	Priority	Type of intersection
Road	Lightly injured	Vehicles involved
City	Day of week	Section of road without intersection
Isle	Road type	Seriously injured
Weather	luminosity	Road work
Province	sidewalk	Type of accident
Zone	Restricted visibility	Traffic Volume
Year	Road network	Specific measures
Month	Autonomous region	Road surface
Hour	Total victims	Grouped zone

Table 2: DGT public features of the vehicle involved in an accident.

Field names	
Type	Vehicle registration year
State	Vehicle registration month
Occupants	Dangerous goods
Year	Burnt car

Table 3: DGT public features of injured people.

Field names		
Age	Gender	Driver’s year license
Position	Security accessories	Adverse effects
Year	manoeuvre	Speed infraction
Pedestrian action	Pedestrian infraction	

Our approach is based on real-time data indexing of traffic incidences from the Infocar DGT website [11]. Data extraction is performed sequentially through real-time web scraping of all registered incidences. After this, all of them are stored in a Cassandra [12] ten node cluster.

The data extracted from Infocar during the previous year has been used in this research. We additionally stored incidences caused by the start and end of a roadwork, geographical location of speed control radars, and of course, the data relating to a recent accident. Table 4 shows some of the general data we are currently indexing from the published incidences. As mentioned, they are stored on our cluster every five minutes.

Table 4: DGT public features of all incidences.

Cause	Type
Subtype	Incidence ID
City	Province
Initial km point	Final km point
Latitude	Longitude
Level (severity)	Description
Timestamps	Lane direction

2.3 Feature extraction

After the preliminary research described before, we now define more complex features that can be derived from the previous simple features from Table 4. As mentioned earlier, these features are new and have not been employed before in traffic studies and allow us to analyse the problem through a new advanced analysis model, Fig. 1.

2.3.1 Case 1: Analysing accidents at section in roadwork

The originality of this research consists in analysing if an accident has happened between two points, initial and final, of a roadwork. For this analysis, all incidences indexed as 'roadwork' are checked with every accident indexed in Cassandra for the same road. We have coded some functions in Python to check this. Of course, timestamps from roadwork and accidents are also checked.

2.3.2 Case 2: Analysing accidents next to a speed control radar

From our personal experience we infer that dangerous situations occur in the proximity of radars due to the sharp slowdown in the chain by multiple vehicles approaching it when driving with a higher speed than allowed.

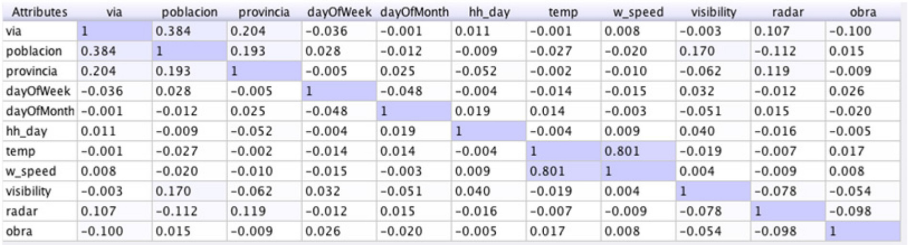
To find out whether the accident is close to a radar, the situation is more complex, because it is necessary to calculate the distance between two points on a sphere. In order to calculate this, first of all we check if the accident happened on the same road where the speed control radar is placed. Then, the distance between them is calculated through the Haversine formula, eqn (1),

$$d = 2r \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos\left(\phi_1\right) \cdot \cos\left(\phi_2\right) \cdot sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{1}$$

where ϕ_1, ϕ_2 and λ_1, λ_2 refer to the latitude and longitude, both expressed in radians, of the two points respectively, and r is equal to the mean radius of the earth. Once we know the distance *d* between the two points, it is checked if this distance is less than 500 meters before or after the speed control radar.

2.3.3 Extracting weather features

Because we have no historical weather data, we obtain it from the W underground API [13]. In order to do this, we post the geographical position and occurrence timestamp of each event. The



Attributes	via	poblacion	provincia	dayOfWeek	dayOfMonth	hh_day	temp	w_speed	visibility	radar	obra
via	1	0.384	0.204	-0.036	-0.001	0.011	-0.001	0.008	-0.003	0.107	-0.100
poblacion	0.384	1	0.193	0.028	-0.012	-0.009	-0.027	-0.020	0.170	-0.112	0.015
provincia	0.204	0.193	1	-0.005	0.025	-0.052	-0.002	-0.010	-0.062	0.119	-0.009
dayOfWeek	-0.036	0.028	-0.005	1	-0.048	-0.004	-0.014	-0.015	0.032	-0.012	0.026
dayOfMonth	-0.001	-0.012	0.025	-0.048	1	0.019	0.014	-0.003	-0.051	0.015	-0.020
hh_day	0.011	-0.009	-0.052	-0.004	0.019	1	-0.004	0.009	0.040	-0.016	-0.005
temp	-0.001	-0.027	-0.002	-0.014	0.014	-0.004	1	0.801	-0.019	-0.007	0.017
w_speed	0.008	-0.020	-0.010	-0.015	-0.003	0.009	0.801	1	0.004	-0.009	0.008
visibility	-0.003	0.170	-0.062	0.032	-0.051	0.040	-0.019	0.004	1	-0.078	-0.054
radar	0.107	-0.112	0.119	-0.012	0.015	-0.016	-0.007	-0.009	-0.078	1	-0.098
obra	-0.100	0.015	-0.009	0.026	-0.020	-0.005	0.017	0.008	-0.054	-0.098	1

Figure 2: Correlation matrix.

mentioned web service returns the data relative to the closest meteorological station, and we are able to add the following interesting data to our feature vector:

- Pressure in mBar
- Temperature in Celsius
- Relative humidity %
- Wind speed in km/h
- Wind direction description (i.e., SW, NNE)
- Visibility
- Precipitation in mm

2.4 Dataset analysed

The dataset used contains a record of 5,157 accidents from May 2015 to January 2016; each table includes data from the road and geographical location of the accident, weather reports, and data relative to the closest speed control radar or its situation inside a section in roadwork.

In addition, we have studied 17,573 incidents during the same period where a roadwork section was found, and 1,247 speed control radars located in the Spanish geography.

3 RESULTS

3.1 Single variable correlations

Our first step in the data analysis is to discard those variables that are not useful to our study. In order to do this, we obtain the matrix correlation shown in Fig. 2. It can be observed that a strong positive correlation exists between wind speed (w_speed) and temperature (temp) with a correlation index (r = 0.384).

The rest of the features are not correlated, making it possible to conclude that they are independent of one another.

3.2 Distribution of the most relevant features

It has been considered interesting to show some of the distributions that have been obtained from our preliminary data analysis. This study has analysed the number of accidents per road extracted from

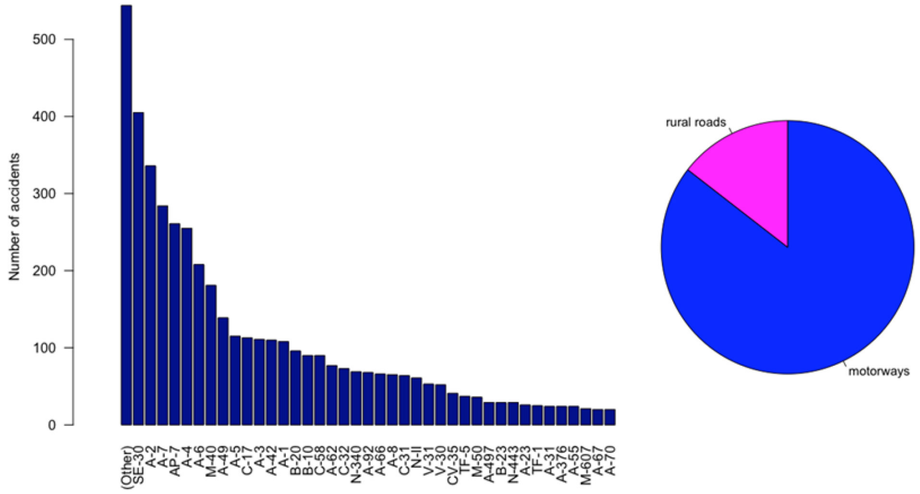


Figure 3: Accident distribution by road and by road type.

the dataset of 5,176 accidents. Next, we present absolute results that classify provinces according to the number of accidents, as well as day of the week and hour of the day.

3.2.1 Accident distribution by road

If we analyse the details on the distribution of accidents by road, (Fig. 3), it is relevant to observe that the SE-30 is the motorway with more accidents, and it is located in Seville, the Spanish province that stands third with regard to rate of accidents. It is followed by A-2 motorway with a 17% less number of accidents, and A-7 that is 29% lesser.

3.2.2 Accident distribution by road type

The number of accidents detected by type of road has been much higher on motorways than on secondary roads, representing 85% compared to a mere 15% respectively (Fig. 3).

3.2.3 Accident distribution by province

When this research was started, it was expected that a greater number of accidents would be found in provinces with higher indexes of population. This has been proved in the provinces of Barcelona and Madrid: both of them with the higher rate of accidents around the Spanish territory (Figs 4 and 5). However, it was very surprising to find that the volume of accidents in Seville, a similar extension province like Valencia and next in the ranking by accident records, reported 50% more accidents than Valencia, next in the ranking by accident records, but with a 50% less accidents than Seville.

3.2.4 Accident distribution by day of week and by hour of the day

Figure 6 represents the accumulated traffic accidents during the whole period studied by the day of the week when they happened, and also represents the distribution of accidents during the hour of the day.

This allows us to define a total *a priori* probability of an accident in any road in Spain given the day of the week. The main inference that we can make is that on Saturdays and Sundays the number

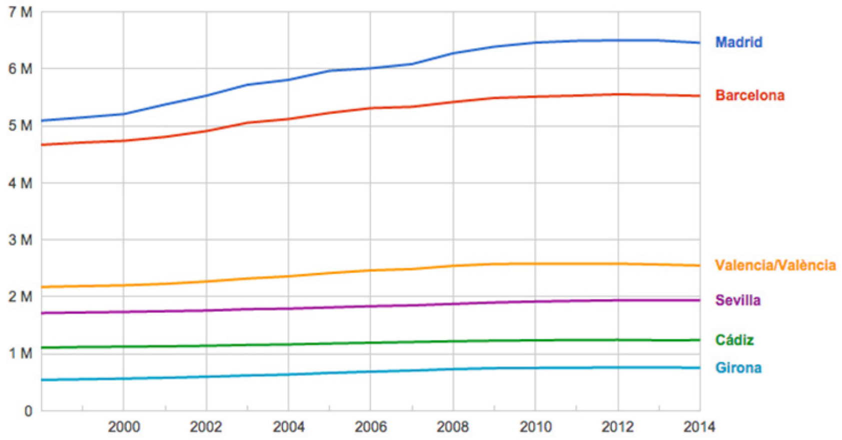


Figure 4: Population by province and year.

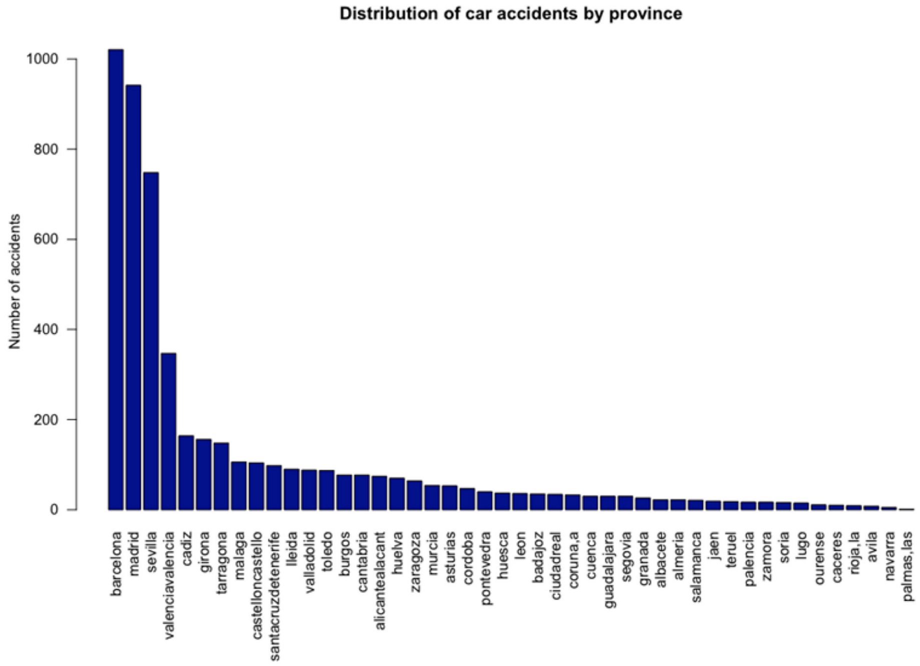


Figure 5: Distribution of car accidents by province.

of traffic accidents is significantly low. It is remarkable that the distribution is similar for the rest of the days but increases on Fridays.

Also, we observe three main peaks at 8:00–9:00 h, 14:00–15:00 h and 19:00–20:00 h. This distribution can be easily explained by the Spanish working hours that generally coincide with these rush hours.

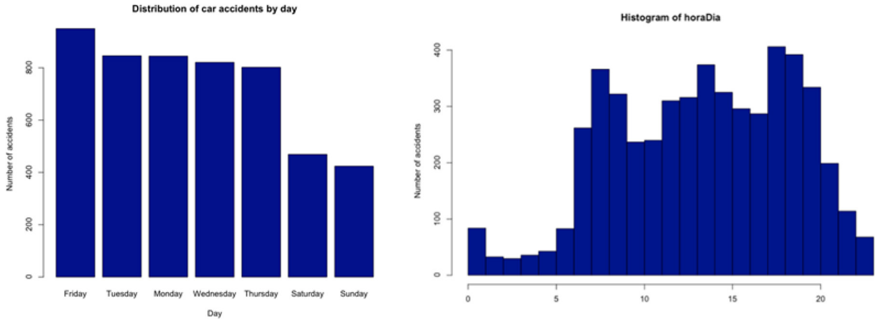


Figure 6: Distribution of car accidents by day of week and by hour of the day.

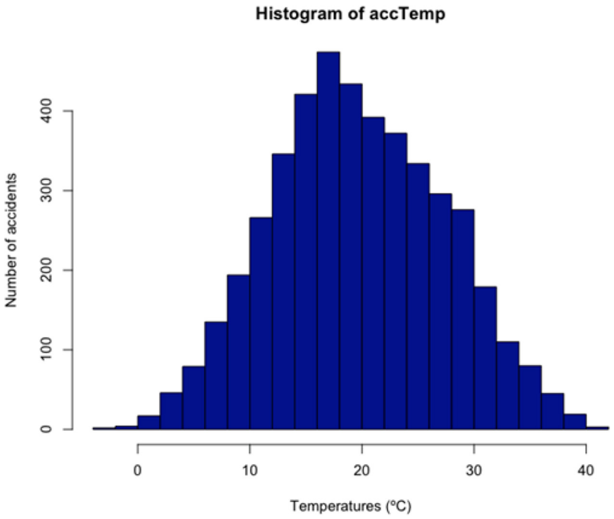


Figure 7: Distribution of car accidents by temperature.

3.2.5 Accident distribution by temperatures

Figure 7 represents the total number of accidents in a histogram of temperatures. This graph presents a distribution that could seem like a Gaussian centered at 19 degrees Celsius. This preliminary result will be used as a starting point to further analyse the dependence of warm weather with the occurrence of accidents.

3.2.6 Number of accident produced near a speed control radar

In this section we present the validity of the speed radar control cameras as a safety measure to reduce traffic accidents by means of adjusting the speed of drivers to the maximum established in the road. Our personal experience suggested that drivers tend to abruptly adjust their speed when suddenly observing these radars, which may lead to risk situations.

In order to analyse these situations, two different distances have been taken (200 and 500 meters) of the accidents to the nearest radar situated in the same road. The accidents that take place inside these radii accounted for the statistics. The main conclusion of the study after processing the whole dataset is that approximately 5,35% of the accidents takes place inside a radius of 500 m of the nearest radar (a total of 276 accidents). When a radius of 200 m of the nearest radar is employed, a 2,87% of accidents occurs (a total of 148).

It can be concluded that it is possible that some speed control measures are not functioning as desired and it is more convenient to perform a study of their location to achieve a true objective.

3.2.7 Number of accidents observed in a section under roadwork

In this article, we also focus on the number of accidents that take place in a section under roadwork. The great number of roads that had sections under renovation or construction suggested the need to find the total number of accidents that took place under these conditions. This study reveals that a total of 30% of accidents occurred at a roadwork section when it was in progress.

4 CONCLUSION

This article presents a study of road traffic accidents on Spanish roads. Each traffic accident has been considered as a discrete and independent random event. In order to study the set of accidents, we have obtained a number of variables associated to each one from several open data sources. Thus, the probability of occurrence of an accident can be modelled by a large number of features, such as weather conditions, incidents caused by the start and end of a roadwork, geographical location of speed control radars, roadway infrastructure, etc. The total number of accidents in the period studied does not constitute a big data problem by itself; however, the large number of variables involved imply that big data techniques must be deployed to extract and store the feature vectors.

In this study, a more complex model is presented that separates from the classical triangle that relates driver, road and accident. On the contrary, a large number of other features are included, which allow an in-depth analysis of the accidents. In addition, two novel features are presented, such as the proximity to speed control radars of accidents and the accidents occurred in roadwork sections. The results have presented in absolute terms the roads in Spain that have had more accidents during the period studied. In addition, a distribution of accidents with respect to temperature has been presented. In addition, the number of accidents with respect to the day of the week has been shown.

An important conclusion of our study is the need for more precise data from official sources that could model the accident. Especially, we have found great problems in the GPS precision of the accidents. The period of study of accidents had to be restricted as well since we had no possible choices to correlate with datasets published in the past.

The results presented in this article will allow us to further focus our research on prediction models that could suggest the probability of occurrence of accidents in roads depending on current variables as well as the need for changes in the infrastructure to prevent them.

ACKNOWLEDGEMENTS

This work has been supported by the Generalitat Valenciana through the project GV/2015/031: Creación de mapas topológicos a partir de la apariencia global de un conjunto de escenas.

REFERENCES

[1] European Union, Smart seat and seatbelt to help sleepy drivers stay alert. *Research*eu Results Magazine,* **42**, pp. 6–7, 2015.

[2] Sabey, B.E. & Taylor, H., The known risks we run: the highway. *Societal Risk Assessment*, ed. R.C. Schwing,. & Albers, W.A., Springer US: Boston, pp. 43–70, 1980.
http://dx.doi.org/10.1007/978-1-4899-0445-4_3

[3] Hakim, S., Shefer, D., Hakkert, A., & Hocherman, I., A critical review of macro models for road accidents. *Accident Analysis and Prevention*, ed. Abdel-Aty, M., Elsevier, pp. 379–400, 1991.

[4] Wang, C., Quddus, M.A. & Ison, S.G., Predicting accident frequency at their security levels and its application in site ranking using a two-stage mixed multivariable model. *Accident Analysis and Prevention*, ed. Abdel-Aty, M., Elsevier, pp. 1979–1990, 2011.

[5] Simoncic, M., A bayesian network model of two-car accidents. *Journal of Transportation and Statistics*, ed. Jeeves, A., United States Bureau of Transportation Statistics, pp. 13–25, 2004.

[6] Davison, P.A., Interrelationships between british drivers' visual Abilities, age and road accidents histories. *Ophthalmic and Physiological Optics*, ed. Oxford Pergamon Press, pp. 195–204, 1995.

[7] Zegeer, C.V., Reinfurt, D., Hummer, J., Herf, L. & Hunter, W., Safety effects of cross-section design for two-lane roads. *Transportation Research Board*, ed. US Department of Transportation, 1987.

[8] Kalokota, K., Seneviratne, P.N., y Center, U.T., *Accident Prediction Models for two-lane Rural Highways*, Utah Transportation Center, 1994.

[9] Shankar, V., Milton, J. & Mannering, F., Modeling frequencies as zero-altered probability processes: an empirical inquiry. *Accident Analysis and Prevention*, ed. Abdel-Aty, M., Elsevier, pp. 829–837, 1997.

[10] DGT Statistical website, available at https://sedeapl.dgt.gob.es/WEB_IEST_CONSULTA/

[11] DGT Infocar, available at http://infocar.dgt.es/etraffic/

[12] Apache Cassandra, available at http://cassandra.apache.org/

[13] Wunderground, available at http://www.wunderground.com/